文章编号：1003-0077(2010)02-0014-10

中文词汇网络：跨语言知识处理基础架构的设计理念与实践

黄居仁[1,2]，谢舒凯[3]，洪嘉馡[4]，陈韵竹[1]，苏依莉[1]，陈永祥[5]，黄胜伟[1]

（1.“中央研究院”语言学研究所，台北；2. 香港理工大学 人文学院，香港；
3. 台湾师范大学 英语学系，台北；4. 台湾大学 语言学研究所，台北；5. 台湾大学 资讯工程学研究所，台北）

摘 要：中文词汇网络(Chinese WordNet，简称 CWN)的设计理念，是在完整的知识系统下兼顾词义与词义关系的精确表达与语言科技应用。中文词义的区分与词义间关系的精确表征必须建立在语言学理论，特别是词汇语义学的基础上。而词义内容与词义关系的发掘与验证，则必须源自实际语料。我们采用的方法是分析与语料结合。结合的方式则除了验证与举例外，主要是在大量语料上平行进行词义标记，以反向回馈验证。完整、强健知识系统的建立，是兼顾知识本体(ontology)的完备规范(formal integrity)和人类语言系统内部的完整知识。我们采用了上层共享知识本体(SUMO)来提供知识的规范系统表征。

关键词：计算机应用；中文信息处理；中文词汇网络；全球词汇网络网格；知识本体；多语处理；跨语言整合
中图分类号：TP391 文献标识码：A

Chinese Wordnet: Design, Implementation, and Application of an Infrastructure for Cross-Lingual Knowledge Processing

Chu-Ren Huang[1,2], Shu-Kai Hsieh[3], Jia-Fei Hong[4],
Yun-Zhu Chen[1], I-Li Su[1], Yong-Xiang Chen[5], Sheng-Wei Huang[1]

(1. Institute of Linguistics, Academia Sinica, Taibei, China;
2. Faculty of Humanities, The Hong Kong Polytechnic University, Hong Kong, China;
3. Department of English, National Taiwan Normal University, Taibei, China;
4. Graduate Institute of Linguistics, National Taiwan University, Taibei, China;
5. Department of Computer Science and Information Engineering, National Taiwan University, Taibei, China)

Abstract: The design criterion of Chinese WordNet (CWN) is to build a complete and robust knowledge system which also embodies a precise expression of semantic relations. Such precise expression for the Chinese sense division and the semantic relations must be based on linguistic theory, esp. lexical semantics. All word sense examples together with the lexical semantic relations in CWN are all attested with corpus data. Our methodology involves first analyzing language data and then combining the analyzed result with corpus by sense tagging to re-examine the accuracy of the analysis. For formal representation and computational application, a complete and robust knowledge system needs to be equipped with the formal integrity of ontology. The Suggested Upper Merged Ontology (SUMO) is adopted for this purpose.

Key words: computer application; Chinese information processing; Chinese WordNet; global Wordnet grid; ontology; multi-language processing; cross-lingual integration

收稿日期：2009-03-23 定稿日期：2009-05-20

作者简介：黄居仁(1958—)，男，研究员，主要研究方向为语言哲学、语言学概论、汉语语法与语意、计算语言学；谢舒凯(1970—)，男，助理教授，主要研究方向为计算语言学、语意学、语言哲学；洪嘉馡(1974—)，女，博士候选人，主要研究方向为语料库语言学、词汇语意学。

1　前言

中文词汇网络(Chinese WordNet，简称CWN)的设计理念，是在完整的知识系统下兼顾词义与词义关系的精确表达与语言技术应用。中文词汇网络，是信息基础建设完善环境中最重要的环节之一。中研院中文词网小组(Chinese WordNet Group)，结合分析详尽的中文词汇词义数据与网络科技的技术，开发了中文词汇网络。

中文词汇网络中文词义的区分与词义间关系的精确表征，建立在语言学理论，特别是词汇语义学的基础上。而词义内容与词义关系的发掘与验证，则源自实际语料。语言内部知识的完整表达，是建立在完整的词义关系系统上，特别是利用“类义词”(Paranym)[1]整合、对比语意关系为主的词汇网络与界定语意场的不同分类系统(Taxonomy)，更以完整标记的跨语言词义关系作为多语知识系统对应的基础。以上的设计理念，使得中文词汇网络不但提供了中文词汇语意深入研究的基本参考数据，更进一步能支持跨语言的知识整合与应用，如全球词网网格(Global WordNet Grid)的建构与生态环保领域的跨语言知识整合[2]。

本研究以中研院语言学研究所中文词汇网络研究小组从2003年以来大量的词汇词义分析研究成果为基础，收录至2009年初的第六版资料，已有8 836个词形，23 670个词义深入分析之synset数据建构于中文词汇网络上，我们以人性化整合查询接口透过因特网呈现，除了提供相关研究人员以及有兴趣的使用者查询检索外，更希望借此系统作为全球词网——多语、跨语言的基础知识架构对应的连结。

本文内容如下：在第二节由本体研究出发，将呈述我们对于中文词汇意义之判定及表达所作的基础研究，说明词义语意面区分的基础与应用、词义判准原则、词义收入与分合操作原则、词义描述规范和词汇语义关系判定原则等。在第三节中，我们将讨论如何将已经分析的数据转成中文词汇知识检索系统的设计概念，包含有SSMS系统和CWN接口，并介绍词义标记系统：人工标记与自动预测，以及中英双语知识本体词网。在第四节里，要探讨的是词汇语意关系表达与预测，重点在词汇语意关系表达与词汇语意关系自动预测。在第五节里，我们将讨论语言知识整合与应用，对于跨语言知识系统的对比与应用作了详尽的探讨，以及讨论两岸词汇对应的比较。第六节是总结。

2　中文词汇意义之判定及表达

2.1　词义与义面区分的基础与应用

词汇语义学相关研究，在理论语言学上表现出来的，是词汇语意学理论与解释能力的蓬勃发展。而在计算语言学方面则是语意导向的研究题目逐渐成为主流。甚至在认知科学的研究中，意义的心理与脑神经处理问题也开始受到重视。在这些潮流中，词网(WordNet)是最重要的共同基础架构，而词义(sense)的区分则是最关键的基本研究议题。

词网是以词义与语意关系为经纬建立的人类语言知识表达基本架构。建构完成的词汇语意网，一方面可以作为语言学研究的素材，另一方面在信息处理上又可以作为自然语言处理以及诸多实际应用的基石。词网里有两项重要的元素，一是以词义为据的词汇分组(即所谓的同义词集(synset))，另一个就是连系词集的语意关系。以同义词集为节点，透过语意关系相互连系，就形成了表征词汇意义及其关系的语意网络。其中，同义词集的建立可说是最基础的工作。建立同义词义，便是把在语境中能表达相同词义的词汇归为一组同义词集，而多义词则分处多组词集，以表示其不同的词义。据此可知，词汇的词义区辨及其同义词的判断与汇集，便成了最根本的工作。然而，虽有实际的需求，词义区辨的原则在学术上却是尚无定论的议题。为了相关工作的进行，本文希望能讨论并建立一组词义区辨的操作原则，一方面能满足一致性与合理性的要求，另一方面又能作为大量中文词汇词义区辨工作上有用的准则。有了一致性的词义判准，语言知识才能有效处理，也才能把语言知识连结到知识本体(ontology)或转换成概念表达。

2.2　词义判准原则

在本文中，“意义”与“词义”二词有特殊的界定。说话者或分析人员根据自身对某一词汇在语境中传达讯息的理解，区分出相同词汇(形)的不同涵义，我们称之为“意义”(meaning)；不同的人可能有不同的区分方式(依据不同的标准或直觉)。进一步，根据适当的标准，判断初步分析的合理性、进行意义的分合、细分等而获致的最后结果，我们称之为“词义

(sense)"。在某些语境下,词可能有受语境影响而改变的意义,人们可以区分出,但这些意义是暂时的,当上下文语境改变时,又会出现不同的相对意义。这样的意义区分,我们称之为"义面"(meaning facet) Ahrens et al.[3]),是中文词网处理文献中所谓"规则化多义"(regular polysemy)的重要创新。如同一段文本中,"报纸"可以指涉阅读的内容,或纸制品的实体。① 以及规则化的动词名物化,都是义面的重要例子。

我们的词义判准建立在五个基础原则上:(一)一义一项、(二)一物一义、(三)一事一义、(四)义不随境迁、(五)义面由观点与语境定义。除了建立理论完整的词义区辨原则外,并同时提供了可以实证的词义区辨操作原则,并对每个原则提供实例。我们借由这个分析原则进行词义区分的工作,并建构工作接口,将中文词网词义区分资料库的内容已全部在线化,中文词网词义区分的资料可直接进入资料库,不用透过机读格式的转档。在本接口上,我们可以进行词汇的查询,词义的新增、修改以及例句和WordNet同义词集的查询和输入。本资料库可有效地管理词汇与词义,并便利于技术报告的整理和编辑。此外,我们也借由Chinese WordShetch提供较为明确的上下文语境的句子,以验证我们分析的词义是具有可靠性的[4] 。

2.3 词义收入与分合操作原则

CWN研究过程,除了建立理论完整的词义区辨原则外,并同时提供了可以实证的词义区辨操作原则;并对每个原则提供语言实例。这个原则是建立在文献(词典)与实证(语料库)上的。以实证为主,文献为辅,两者有冲突时,则由语言学专业人员分析解决。

词义分和操作的原则,大致上,我们分为三大类:

(一)词义的判定

如何判定两个分析出来的意义确实属于不同词义(应分列不同义项)?或为同一词义的两个义面?判定的标准是不同的词义不会出现在同一语境之下。在实际操作上,可以尝试找出同时带有两个分析出的意义的语境(句子)。若成功,则为单一词义,双义面。若失败,则属两个词义。

(二)"歧义句"与"同时带有两项意义"的判定

有些词汇可以造出歧义句使得语句中的词汇能作两种以上的不同解释,即分析出两个以上的意义。但分析出的多重意义,在语境提供足够讯息后,只能有单一意义留存(即所谓排歧)。所以"歧义句"的情形不能当作该例句同时带有两项意义。例如:"看病"一词,可能是医生替病人治病,表示诊治;也可能是病人接受医生诊治,表示就诊。但是在[他这个医生不看病,只做研究。]的句中,[就诊]的词意被消去了不会出现,这就是典型的歧义与语境消歧。

(三)义面的判定

如何判定两个分析出来的意义确实属于同一词义下的不同义面?判定的标准为,同一词义下的不同义面,会有分属不同义面的语料,但也同时带有两个义面的语料。在实际操作上,则必须让下列两个条件并存。

a. 找出分别反映不同义面的语料。

b. 找出同时反映不同义面的语料。

另外,请注意动词的"名物化"现象,当名物用法同时指涉同一事件的名称时,也析分成不同义面,不作词义上的区分。但是如果名物用法的指涉改变,如转为指涉参与事件的对象或结果,则为不同词义。例如"拨款"一词,词义1的动词表示"支付或调配款项"这个动作;词义2的名词则是指"支付或调配的金钱"。

2.4 词义描述规范

黄居仁等[5] 提出的词义区辨原则与操作原则,是中文词义数据库建立文件与《词义区辨小词典》编纂的依据。《词义区辨小词典》所收录的词条(entry),以现代汉语通用语词为范围,不列入现今已不用或罕用的词汇。而收录的中文词汇条目,包含单字词、双字词和多字词。本词典尽可能提供各词目(lemma)完整而且正确的讯息,包含标音(汉语拼音与国语注音)、释义、英文对译、词类、例句、附注,如图1所示。

2.5 中文词汇语义关系判定原则

在英语及其他的欧洲语言里,词汇语意关系已有相当充分的研究。例如,欧语词网(EuroWordNet, Vossen 1998)[6] 就是一个以语意关系来勾勒词汇词义的数据库。也就是说,词汇意义的掌握是通过与其他词汇语意的关连来获得的。为了确保数

① 例句如:"今天的报纸登的是有趣的新闻,过期的报纸可论斤回收贩卖;但也有人专爱在旧报纸里找漏网新闻。"报纸在整段中式连贯的主题,理论上必须有相同的词义(sense),但又有不同的意义,故以"义面"区分之。

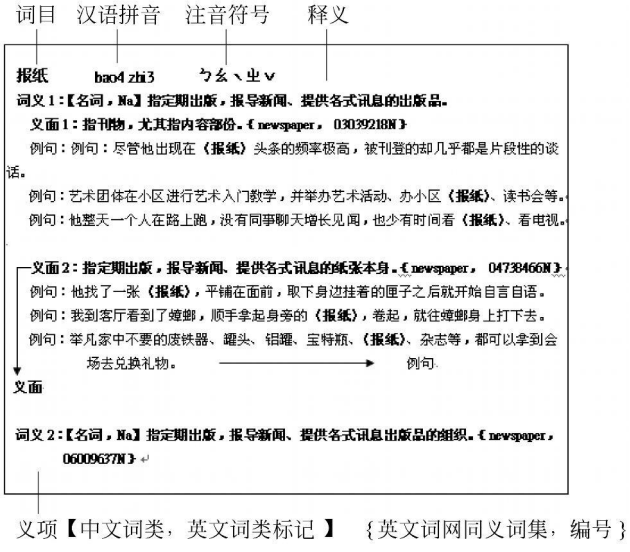

图 1　中文词汇条目内容范例

据库建立的质量与一致性，欧语词网计划就每一个处理的语言其词汇间的词义关系是否成立提出相应的语言测试。实际经验显示，利用这些语言测试，人们可以更容易且更一致地辨识是否一对词义之间确实具有某种词义关系。而且，每一个使用数据库的人也可以检验其中关系连结的正确性。换句话说，对一个可检验且独立于语言的词汇语意学理论而言，这些测试提供了一个基石[7]。

由于词网是一个以词汇间的语意关系为主要收纳对象的数据库，为了确保数据库的质量与一致性以及实际应用的可靠性，一个可以检验语意关系连结适当与否的方法是必要的。就理论的层面来看，这也是为词汇语意学建立实证基础的第一步。为探究中文词义关系建立中文语言测试的可能性，我们尝试为一些重要的语意关系提供测试的句式和规则来评估其可行性。这项研究除了建构中文词汇语意学的理论基础，也对 Miller 的词汇网络架构[8]提供了一个有力的支持，这个架构在词汇表征和语言本体架构研究上开拓了关系为本的研究思路[7]。

3　中文词汇知识检索系统设计

CWN 至 2009 年初，目前累积的成果，共有超过 8 700 多个词形，23 000 多个词义，平均一个词形约有 2.67 个词义；词义数在两个以上的词形，约有 4 500 个，其中，“打”这个词形，具有最多的词义，共 125 个。在这些词义中，动词的词义约 10 300 个，名词的词义约 10 400 个；义面的数量约有 5 500 个。这个大规模数据库的维护、更新，以及提供相关研究人员查询使用，都需要有效率的工作平台，中文词汇知识检索系统之开发，在上述动机下完成。除了研究成果共享之目的外，更希望借此作为中文词汇知识网络研究之基础架构。

3.1　SSMS 系统

为了可以让机器读取并储存大量的词汇词义区分的数据，我们以词汇知识为基础，来整合词汇词义的信息，开发了中研院词汇词义管理系统（Sinica Sense Management System），简称 SSMS[9]。在 SSMS 里，包含了中文词网小组所收录并分析的词条、词义等相关信息。换言之，SSMS 包含的词条信息有：词类、例句、对应 WordNet 的英文同义词集（synset）、词汇语意关系如：同义词、反义词、上位词、下位词等等。

从 2004 年 2 月起，中文词网词义区分数据库的内容已全部在线化，中文词网词义区分的数据可直接进入数据库。在该管理系统下，我们可以进行词汇的查询，词义的新增、修改以及例句和 WordNet 同义词集的查询和输入。本数据库可有效地管理词汇与词义，并便利于技术报告的整理和编辑。

3.2　CWN 界面

中文词汇知识检索系统在设计阶段考虑了使用者角度与系统功能发展角度，共同建立起系统架构与操作流程，详细描述系统范围内相关数据结构和操作步骤，特别是设计一套整合式实时查询的方式[10]，提供系统使用者一个整合查询接口快速查询以及浏览有兴趣的各个词义信息。系统提供的查询范围，有：中文词汇、释义内文、英文对译、中文词汇模糊查询、注音、汉语拼音等，使用者可依不同信息或不同需求来选择查询的方式。主要的出发点是能对词汇与语义相关连的内容，做广泛而有效的检索，也是借着检索的比对，来确保释义语言及语义区分的一致性及强健性。查询结果以词汇编号为主键由数据库中提取出词目、词义、领域、释义、语义关系、英文对译、例句及附注等项目依序排列，通过浏览器可清楚呈现给使用者。

我们于 2006 年，将中文词网词义区分数据库的成果网络化，以便提供给使用者直接查询。目前，我们命名为“中文词汇网络（Chinese WordNet），http://cwn.ling.sinica.edu.tw/”[11]。

3.3 词义标记系统

3.3.1 人工标记

为了彻底呈现语言的真实性，我们对于每一个词条(lemma)的词义(sense)和义面(meaning facet)做了详尽的区分，同时，也借由这些词义和义面，开发出一套词义标记系统[12-13]。中文词汇网络的词义，基准，承袭了普林斯顿词网的既成传统，以同义词集为节点，透过语意关系相互联系。建立同义词义，便是把在语境中能表达相同词义的词汇归为一组词集，而多义词则分处多组词集，以表示其不同的词义。为了证实我们所分析的词义可以完整地表现在实际语言上，我们开发了设计出一个超过 11 万词的大规模中文词义全文标示语料库，以我们已经分析过的词义作为基础，以中研院平衡语料库为标示对象，从中摘录 56 篇完整文章，利用 N-gram 与搭配信息等语言知识，结合机器学习方法作为自动词义标示的预处理工作，然后将自动标示结果进行人工校正。

3.3.2 自动预测

大量精确的词义标注资料，可提供多项计算语言相关研究的丰富素材。但是，中文语料库词义标记主要的瓶颈是缺乏自动标记参考的资料，而人工标示成本昂贵，造成语料库语意标示工作的困难。近年来的许多研究，显示出对大规模词义标示语料集的大量需求，这些资源在建构上是否完备，往往会影响整个研究进行方向以及研究结果的正确性。为了克服此问题，本研究发展一套半自动词义标示方法，作为标示词义的预处理，再经由语言学专业人士校订。语料库制作以中研院平衡语料库为对象，从中摘录文章，并对摘录出的文章中的词做词义标示，形成一个大规模的中文词义标示语料集以供自然语言处理研究使用[12-13]。

根据柯淑津等[12]的研究，自动标示词义的方法，采用诱导式方法（Bootstrap）逐步放宽标示条件，来扩增标示语料，其系统组织如图 2 所示。

自动标示词义的第一阶段采用 N-gram 模式，将标示出词义的资料加入训练集中，以作为第二阶段的训练语料。利用 N-gram 处理词义标示是基于下面的假设：存在包围目标词前后 N 个词完全相同的两个子句，我们推论它们应拥有一样的词义。在此使用 N-gram 有两项主要目的，第一是扩大训练集，因语料库中常可见到相似之子句。第二个目的是过滤训练资料集的噪声，以此检验人工标示资料之不一致性。第二个阶段我们使用搭配信息来增加标示集数量，搭配信息是一种很强的语言关系，能决定目标词汇之词义。利用其所搭配的共现词汇(collocation)与词义特征(semantic feature)等进行词义的预测，第一部分词义标示以词义下再细分至义面为准，结果如表 5 所示，其整体的正确率为 57.47%。第二部分我们将词义标示处理至词义为止，不再细分义面，其整体的正确率大约可提升至 64.51%。

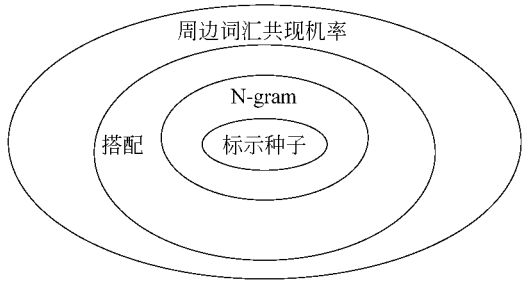

图 2 标示词义系统组织图

3.4 中英双语知识本体词网

为了追求语言知识架构的丰富性，我们采用“建议上层共享知识本体”(Suggested Upper Merged Ontology，简称 SUMO)[14] 为基础来进行语言知识的对照。在自然语言中，常会有一词多义的现象，甚至于有模棱两可的现象，诸如此类的词汇，并不能以一般的词义就能分析的精准，因此，在我们的系统里，还引入了义面(Facet)的概念，不但使我们的系统有更强的表达能力，更使我们能够描述某些随语境转变成共现的语义区分，同时，我们利用类义词来表达词义间的同类聚合关系[15]。

中研院中英双语知识本体词网(The Academia Sinica Bilingual Ontological WordNet，简称 Sinica BOW)[16] 以 WordNet[17] 为基础，加入中国台湾地区所使用的中文经验，搭配领域以及 SUMO，并以 WordNet 的 1.6 版和 1.7.1 版之名、动词的单词义和多词义对应资料为基础，作为媒介连结了领域词汇库和领域知识本体。我们将已经分析完成的 Chinese WordNet 的词义，拿来对应 WordNet1.6 的 synset（总共有 99 642 个数据），发现可以对应上的总共有 18 055 个词义，目前的涵盖率(coverage)为 18.12%。此外，系统借由多元、友善的接口，将功能切割为词网、知识本体以及索引三个主要单元，提供跨语言信息转换、词义的区分与词义关系的连结、语言信息与概念架构(知识本体)的连结以及使

用领域等信息。

Sinica BOW[18] 主要使用的资源包含 WordNet、ECTEC(English- Chinese Translation Equivalents Database)以及 SUMO。ECTEC 是以 WordNet 为基础，经由现有英中或中英电子辞典的词形对应，替每个同义词集的词义找出可能相对应的中译词组，再经由人工检验。寻找对译的过程中，尽可能的以词汇而非描述性短语表达，目的在于让每个同义词集都有最适当的一至三个左右的中文对译。

SUMO 则是由 IEEE 标准上层知识本体工作小组所建置，其目的在于促使自然语言处理、信息检索、自动推论以及资料互通性等工作的进行。知识本体类似于字典或词汇表，但讯息更丰富，以便于计算机处理其内容。知识本体以格式化的方式表达概念(Concept)、关系(relation)以及公理(axioms)。上层知识本体是将一般性、后设性(meta)、摘要性以及哲学类的概念指出，所以特殊领域的概念可由其中的概念所涵盖，但特殊领域概念的知识本体则期许由各领域自行制订[19-20]。日前 SUMO 已经与 WordNet 1.6以及 2.0 版本结合，且以同义(synonymy)、上位(hypernym)、体例(instantiation)这三种类别显示同义词集和 SUMO 概念间的对应关系，例如：同义词集 cell(细胞)与细胞概念(cell)是同义。Hockey(曲棍球)属于运动概念(sport)，两者间的关系为上位，也就是说运动涵盖 hockey(曲棍球)。China(中国)属于国家(nation)这概念的体例。我们利用 Sinica BOW 的系统(http://bow.sinica.edu.tw/)查询 WordNet 的词汇，得到的相关讯息与 SUMO 的讯息。

4　词汇语意关系表达与预测

4.1　词汇语意关系表达

在英语及其他的欧洲语言里，词汇语意关系已有相当充分的研究。欧语词网[6]就是一个以语意关系来勾勒词汇词义的跨语言词汇知识库。也就是说，词汇意义的掌握是通过与其他词汇语意的关连来获致的。为了确保数据库建立的质量与一致性，欧语词网计划就每一个处理的语言其词汇间的词义关系是否成立提出相应的语言测试。实际经验显示，利用这些语言测试，人们可以更容易且更一致地辨识一对词义之间是否确实具有某种词义关系。而且，每一个使用数据库的人也可以根据测试检验其中关系连结的正确性。换句话说，对一个可检验且独立于语言的词汇语意学理论而言，这些测试提供了一个基石。

在 CWN 的语意关系标记选择上，我们除了参考了普林斯顿 WordNet 的语意关系连结，例如：同义词、反义词、上位词、下位词……等等，另外，我们也开发了“类义词(paranymy)”[1,15]的语意关系连结，主要是以 WordNet 为框架，在姊妹词汇(Sister Terms)中解释丰富的概念关系(Rich Conceptual Relations)。根据黄等[1]之前的研究，以词汇语意关系的角度来定义这些姊妹词汇，我们就称之为“类义词”。增加“类义词”的运用，可以让我们在词网的描述上更具完整性，也更丰富了词汇的的本体知识。完整的“类义词”语意关系会依照同一个概念原则来将他们归类为同一类，我们深深地相信，我们将“类义词”当作词汇的语意关系，对于我们在处理词汇语意关系上，有更明确的描述与对应，也可以借此提供更明确的讯息以及富有 ontological 的词网。

根据黄居仁等[1]的分析，我们将“类义词”分为两大类：一、相对类义词(Contrary Paranymy)；二、重迭类义词(Overlapping Paranymy)。“相对类义词”，通常是有比较级和最高级的；可以用 very，almost 等词修饰的词汇，这个语意关系的词汇也可以是中等程度的词汇，所以在描述上有可能既不是这个也不是那个。例如：某个东西可能被认定是“温的”，因为它既不是热的也不是冷的。除此之外，通常“相对类义词”又可被分类成“认知、感官类”(perceptional paradigms)或“约定俗成类”(conventional Paradigms)。“认知、感官类”是基于人类的认知、感知与感官，例如：快/慢的上层节点是速度。至于速度是快或是慢，这个就完全依靠个人的感知，然而这样的感知是不同于其他人的。“约定俗成类”就如同在中文里，我们对于双亲这个概念所使用的称呼词汇，我们可以图 3 来表示这样的关系。

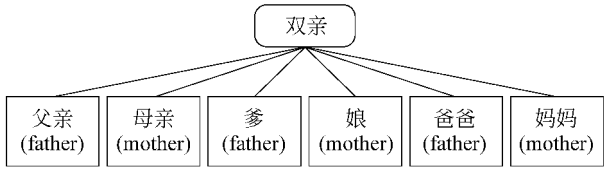

图 3　双亲称谓示意图

至于“重迭类义词”，简单来说，则是这个类型的两个姊妹词汇共同拥有一些相同的特征，在 CWN 里面，我们基于语言约定俗成的用法来做解释与区

分这类重迭类义词，通常与语言的使用与经验相符合。例如："箱子"和"盒子"，所隐含的语意与概念是相同的，但是，当我们要装置较大的物品，如：电视机或计算机的时候，我们会使用"箱子"；当我们只是要放置较小的物品，如：杯子或饼干，我们就会使用"盒子"。又例如："警官"与"警员"，位阶较高、权力较大的，我们称之为"警官"；反之，则称之为"警员"。尽管如此，"警官"与"警员"，所呈现的语意概念，在某些方面是相同的。

4.2 词汇语意关系自动预测

如前所述，词汇网络是以"同义词集"(Synset)与词汇语意关系为核心所架构出之词汇知识系统。也就是说，词汇网络架构表达的不仅是词汇本身的概念性知识，它亦表达了词汇之间的语意关系。然而，从普林斯顿英语词网以及欧语词网的建构经验来看，这是一项相当费时耗力的庞大语言工程。对于经费取得困难、使用频度较低之语言而言，建立此项语言资源更为不易。从词汇语意与知识表达的角度观察，我们认为不同语言对于概念原素(Conceptual Atoms)，可能有着不同之表达方式，但是在词汇语意关系的表达上，则应具有更大程度之"普同性"。因此利用"诱导式方式"(Bootstrapping)针对已发展成熟之英语、欧语词网之语意关系，以加速新的词网雏形成形，就成了一个自然而然的另类选择。

在上述背景之下，我们提出"多语词网诱导"。词汇语意关系自动标记模型(Bootstrapping from Multilingual Wordnets)。此模型是基于中文词汇网络小组一系列之相关研究所得出[5, 9, 21]。我们假定在词汇语意关系之标记上，可以借力于其他已成形之词网的跨语言词义关系资源。主要提出的论点在于，利用词汇关系隐含的代数特性(如及物性等)，转化成进行词义标记所涉及之逻辑条件，并以反向回馈验证。在文献[5]中，曾针对 210 个中文词形(Lemma)做过小规模之试验与评价。以此为基础[21]，进一步在规模上与多语扩充两个面向上作延伸试验。亦即，在规模上，我们将目前中文词网小组所定义完成之 8 000 多笔中文同义词集纳入；在多语扩充上，我们将欧语词网[6]亦纳入实验对象。其中包括了德语、法语、捷克语、荷兰语、西班牙语、义语与爱沙尼亚语等七种欧洲语言。实验结果相当具有前景，我们相信此模式可以协助一个词网的快速原型(rapid prototyping)发展，加速词网核心资源的建构过程。

5 语言知识整合与应用

着眼于作为一个跨语言知识处理架构，中文词汇网络的发展过程中，亦与欧洲语言(法语、意大利语)、日语以及两岸中文之词汇对应进行了语言知识整合与应用之尝试。

5.1 跨语言知识系统的对比与应用

在跨语言知识资源平台设计上，中文词汇网络小组与意大利国家计算语言学研究所，针对跨语词汇知识资源之协作机制，共同提出了称之为 Lex-Flow 之分布式计算架构，并以意大利文与中文为例，成功地展示了初步之实验成果[22-23]；在跨语词汇知识表达之形式化上，我们亦与法国土鲁斯大学合作，利用图形处理与语意空间邻近度计算技术，建构了中法动词语意对应网络[24]，同时亦通过心理语言学实验得到心理处理历程之初步验证。此外，以 WordNet 的讯息作为中介，我们比对了中、日文之汉字知识表征之相同与差异[25]。

为了解决全球多语化所带来的问题，我们需要一个跨语言的知识信息整合平台。我们需要一个知识信息系统，其设计之核心，在于产生内容可协作的(content interoperability)标准化制作、跨语言之分散性知识资源共享与交换机制，及其存取与检索接口。在工作方法上，我们将以知识本体驱动的方式，利用上层知识本体与全球词汇网络网格之串接作为知识资源核心，并辅以文本知识发掘与语意分析技术(请参见 Kyoto 计划，文献[2])。以此方法所产生的知识资源，将以 wiki 平台之呈现，使其得以通过相关领域专家与使用者之回馈得到维护与永续性。最后，我们亦将会拓展此模式到不同语言与文化领域。此跨国合作计划之最后目标，即在于设计与实现这样的系统与资源，对于巨量之分散于全球的知识，可以用一致的格式加以表达呈现，并进行深层之概念检索与发掘。此项研究成果，特别是对于全球之中小型企业，包括非营利性组织，将有相当大之帮助。

5.2 两岸词汇对应

在我们的中文词汇网络里，我们结合了 WordNet 的讯息，利用各种不同的语义关系，将每个原本属于看似独立的词义连结起来；也以 WordNet 为中介，比对同样是中文而区分出来繁体中文

系统(Chinese WordNet，CWN)与简体中文系统(Chinese Concept Dictionary，CCD)[26-27]。

自从2000年开始，北京大学计算语言学研究所就已经开始着手以WordNet为基准，研究CCD，并建立一个中英双语的词网，一个可以提供各种不同研究的词网，如机器翻译(MT)，信息检索(IE)…等等。

CCD，中文概念辞典，是一个中英双语的词网，由北京大学计算语言学研究所开发，整个架构发展也是来自于WordNet[28-30]。在CCD的发展手册里记载，研究团队描述这些词义的首要条件，是不可以破坏原本WordNet对于同义词集定义概念与其语义关系的架构。另一方面，CCD的研究团队考虑到可以存在许多在中文与英文的不同描述架构，所以，他们不止表现对中文词汇内涵的表达，也发展了中文词汇语义与概念的关系性，以利于强调中文的特质。

CCD的研究团队专注在整个CCD的架构，提出同一概念的同义词集的定义，其所呈现的概念、定义和概念网的上下位语义关系，每一个同义词集都有其基本关系，彼此之间亦有语义关系的存在。至于CCD的逻辑推演原则在语义网上的呈现，是运用到数学的形式而来的，是可以帮助研究者在中文语义分析上的使用。CCD的总体结构沿用WordNet框架：CCD以同义词集(Synset)定义概念(Concept)，在概念之间定义关系(Relation)；涉及词性有名词、动词、形容词和副词，主要的关系有同义关系、反义关系、下位关系、整体部分关系和词法关系等[29]。

繁体中文系统的英中对译(CWN)与简体中文系统的英中对译(CCD)，依不同词类，区分成：名词、动词、形容词和副词四大类来进行对比，以WordNet为主，检测在同一个Synset中，繁体中文系统的对译词汇和简体中文系统的对译词汇，然后再进行比对。

在四大词类中，我们可以清楚得知，在同一个Synset中，繁体中文系统，可能有多个相对应的对译词汇，同样地，简体中文系统也可能有个相对应的对译词汇。在这些对译词汇里，又有可能是两边使用的对译词汇完全一样，称之“完全相同”；如果，两边使用的对译词汇，没有一个相同的，称之“完全不同”，也就是“真正不同”；或者，只有使用其中一个或一个以上对译词汇，这个状况，称之“部分相同”，而在“部分相同”的对译词汇，如果两边的对译词汇使用的词首相同，称之“词首相同”，如果只是使用到相同的字，则称之“部分字符相同”，详情见表1。

表1　CCD和CWN对译的各种分布状况

Synset	CCD对译词汇	CWN对译词汇	比例	
bookshelf	书架、书柜、书橱	书架、书柜、书橱	60 176 (60.39%)	完全相同
lay off	下岗	解雇	6 762 (6.79%)	完全不同
immediately	立即	立刻	10 628 (10.67%)	词首相同
according	据报	根据	22 076 (22.16%)	部分字符相同
总计			99 642 (100%)	

CCD与CWN的对比研究，以及CWN与欧语词网的对比研究，很清楚指出一个在词义与词义关系的架构上，对两个词汇库系统做宏观研究的新途径；这是词汇语义学拓展研究的绝佳契机。

6　结论

本文的研究，以中文词汇网络建构为开端，以WordNet的信息为桥梁，透过各种语义关系，串连起信息丰富的多语语料，比对同样是中文而区分出来繁体中文系统与简体中文系统这两个系统；也比对了中、日两种不同语言在表达同一概念的相同与差异。当然，因全球多语化的驱动，我们也试图开发出一个跨语言的整合知识讯息平台，并结合上层知识本体与全球词汇网络网格之串接作为知识资源核心，进而发掘语义分析的技术。

中文词汇网络兼顾词义与词义关系，以及在完整的知识系统下的精确表达中文词汇知识的设计理念，为中文与跨语言知识工程与语言科技应用，提供了一个极具价值的基础架构。并且在词汇语义学的基础研究上，以及中文词义标记语料库上，都跨出了重要的一步。更重要的，CWN提供的架构基础，使中文词网能参与欧盟重要国际计划的协作更凸显了不受限于语言的知识架构，在未来语言科技发展中的关键性。

参考文献

[1] Huang, Chu-Ren, I-Li Su, Pei-Yi Hsiao, and Xiu-Ling Ke. Paranyms, Co-Hyponyms and Antonyms: Representing Semantic Fields with Lexical Semantic Relations. [C]//Chinese Lexical Semantics Workshop. May 20-23. Hong Kong: Hong Kong Polytechnic University. 2007: 66-72.

[2] Vossen, Piek, Eneko Agirre, Nicoletta Calzolari, Christiane Fellbaum, Shu-Kai Hsieh, Chu-Ren Huang, Hitoshi Isahara, Kyoko Kanzaki, Andrea Marchetti, Monica Monachini, Federico Neri, Remo Raffaelli, German Rigau, Maurizion Tesconi and Joop VanGent. KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures[C]//To be presented at the 4th Global WordNet Conference. Szeged, Hungary. January 2008. 22-25.

[3] Ahrens, Kathleen, Li-li Chang, Keh-jiann Chen, and Chu-Ren Huang. Meaning Representation and Meaning Instantiation for Chinese Nominals[J]. Computational Linguistics and Chinese Language Processing. 1998, 3(1): 45-60.

[4] Hong, Jia-Fei, Chu-Ren Huang and Kathleen Ahrens. Event Selection and Coercion of Two Verbs of Ingestion[C]//Proceedings of Chinese Lexical Semantics Workshop. 2007: 59-65.

[5] Huang, Chu-Ren, Elanna I. J. Tseng, Dylan B. S. Tsai and Brian Murphy. Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations[J]. Language and Linguistics. 2003. 4(3). 509-532.

[6] Vossen, Piek. (ed.). EuroWordNet[EB/OL]. 1998. Dordrecht, Holland: Kluwer.

[7] 蔡柏生, 黄居仁, 曾淑娟, 林贞仪, 陈克健, 庄元一. 中文词义关系的定义与判定原则[J]. 中文信息学报. 2002, 16(4): 21-31.

[8] Fellbaum, Christiane. (ed.). WordNet: An Electronic Lexical Database[M]. MIT 1998. Press.

[9] Huang, Chu-Ren, Chun-Ling Chen, Cui-Xia Weng, Hsiang-Ping Lee, Yong-Xiang Chen and Keh-jiann Chen. The Sinica Sense Management System: Design and Implementation[J]. Computational Linguistics and Chinese Language Processing. 2005, 10(4): 417-430.

[10] 陈永祥, 洪嘉馡, 黄丽婉, 黄居仁. 因特网中文词汇知识检索系统之建置[C]// 第七届汉语词汇语义学研讨会(CLSW-7). 2006. 台北. 2006. 5. 22-24.

[11] Chinese WordNet[EB/OL]. http://cwn.ling.sinica.edu.tw.

[12] Ker, Shu-Jin, Chu-Ren Huang, Jia-Fei Hong, Shi-Yin Liu, Hui-Ling Jian and I-Li Su. 中文词义全文标记语料库之设计与雏形制作[C]//The 19th ROCLING Conference. Taipei. 2007: 335-346.

[13] Ker, Sue-Jin, Chu-Ren Huang, Jia-Fei Hong, Shi-Yin Liu, Hui-Ling Jian, I-Li Su and Shu-Kai Hsieh. Design and Prototype of a Large-scale and Fully Sense-tagged Corpus[J]. The Third International Conference on Large-scale Knowledge Resources (LKR2008). 2008. Tokyo. March 3-5.

[14] SUMO[EB/OL]. http://www.ontologyportal.org.

[15] Huang, Chu-Ren, I-Li Su, Pei-Yi Hsiao, Xiu-Ling Ke. Paranymy: Enriching Ontological Knowledge in WordNets[C]//The 4th Global WordNet Conference. 2008: 220-228. Szeged, Hungary. January 22-25.

[16] Sinica BOW [EB/OL]. http://BOW.sinica.edu.tw.

[17] Princeton WordNet[EB/OL]. http://wordnet.princeton.edu/.

[18] Huang, Chu-Ren, Ru-Yng Chang, Shiang-Bin Lee. Sinica BOW (Bilingual Ontological Wordnet): Integration of bilingual wordnet and SUMO[C]// The 4th International Conference on Language Resources and Evaluation (LREC2004). 2004. Lisbon. 26-28 May, 2004.

[19] Niles, Ian and Adam Pease. Toward a Standard Upper Ontology[C]//Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19, 2001: 2-9.

[20] Niles, Ian and Adam Pease. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology[C]//Proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003). Las Vegas, Nevada, June 23-26, 2003: 412-416.

[21] Hsieh, Shu-Kai, Simon Petr and Chu-Ren Huang. 大规模词汇语意关系自动标记之初步研究: 以中文词网(Chinese Wordnet)为例[C]//计算语言学国际会议, 新竹, 2006.

[22] Bertagna, Francesca, Monica Monachini, Claudia Soria, Nicoletta Calzolari, Chu-Ren Huang, Shu-Kai Hsieh, Andrea Marchetti, and Maurizio Tesconi. 2007. Fostering intercultural collaboration: A Web Service Architecture for Cross-Fertilization of Distributed Wordnets[J]. T. Ishida, S. R. Fussell, and P. T. J. M. Vossen (Eds.): IWIC 2007, LNCS 4568, 2007: 146-158. Springer-Verlag Berlin Heidelberg.

[23] Soria, Claudia, Monica Monachini, Fransesca Bertagna, Nicolleta Calzolari, Chu-Ren Huang, Shu-Kai Hsieh, Andrea Marchetti and Maurizio Tesconi. Exploring interoperability of language resources: the

case of cross-lingual semi-automatic enrichment of wordnets[J]. Language Resource and Evaluation (Special issue: Multilingual Language Resources and Interoperability). Springer Verlag. 2009: 87-96.

[24] Gaume, Bruno, Laurent Pr vot, Chu-Ren Haung, Shu-Kai Hsieh and Chao-Jan Chen. Building and Aligning Chinese and French Paradigmatic Graphs [C]//CIL18. 2008. Seoul: Korea.

[25] Huang, Chu-Ren, Chiyo Hotani, Tzu-Yi Kuo, I-Li Su and Shu-Kai Hsieh. Wordnet-anchored Comparison of Chinese-Japanese kanji Word[C]//The fourth Global WordNet Conference. 2008 (b). Hungary: Szeged. January 22-25.

[26] Hong, Jia-Fei, Chu-Ren Huang and Ming-Wei Xu. 以中文十亿词语料库为基础之两岸词汇对比研究[C]//第十九届自然语言与语音处理研讨会(ROCLING'07). 台北. 2007: 6-7. 2007: 287-302.

[27] Hong, Jia-Fei, Chu-Ren Huang, and Yang Liu. WordNet Based Comparison of Language Variation: A study based on CCD and CWN[C]//Proceedings of the Third International WordNet Conference. 2006: 61-68. Jeju. Januaray 22-25.

[28] 于江生，俞士汶. 中文概念词典的结构[J]. 中文信息学报，2004，16(4)：12-21.

[29] 于江生，刘扬，俞士汶. 中文概念词典规格说明[J]. Journal of Chinese language and Computing. 2003，13(2)：177-194.

[30] 刘扬，俞士汶，于江生. CCD 语义知识库的构造研究[C]//2003 中国计算机大会，2003.

书　　讯

2009 年《中文信息学报》合订本已出，还有少量过刊合订本，详细定价如下：

出版年	定价/元	出版年	定价/元
1997	30	2004	70
1998	30	2005	70
1999	55	2006	85
2000	55	2007	100
2001	55	2008	100
2002	55	2009	105
2003	55		

愿购者(邮购需加 15%的邮资费)，请按以下地址汇款：

邮编：100190　　通信地址：北京 8718 信箱《中文信息学报》编辑部

电话：010-62562916　　E-mail：cips@iscas.ac.cn