“计算语言学概论”课程知识要点

詹卫东 北京大学中文系

序号	内容提纲	知识要点	进一步阅读文献
第 1 讲 什么是计算语言学	· 计算语言学的研究内容 · 计算语言学的研究方式 · 计算语言学的应用领域 · 计算语言学的发展简史	· 图灵测试（A.M.Turing） · Eliza 人机对话系统（J.Weizenbaum） · 中文屋子思想实验（J.Searle） · 自然语言 vs. 人工语言 · 算法及其性质 · 知识表示、知识获取、知识应用	Warren Weaver (1949) Turing (1950) ALPAC (1964-1966) Shuly Wintner（2009） Kenneth Church（2011）
第 2 讲 语言知识的形式化表达	· 语言能力与语言知识 · 有限状态自动机/正则文法（线性结构） · 上下文无关文法（树结构） · 特征结构与合一运算（图/网结构）	· 对象语言与元语言 · 终结符与非终结符 · 产生式（重写）规则、推导、句型、句子、乔姆斯基范式 · 特征结构（“特征：值”矩阵 AVM） · 特征值嵌套、特征值共享 · 特征结构的包孕关系 · 合一运算、合一的作用	Noam Chomsky (1959) 冯志伟等译（2005）第 1 章、第 10.3.2、11.1-11.3 节，第 13 章。
第 3 讲 汉语短语结构语法体系	· 汉语的语法范畴体系 · X-Bar 理论模型 · 基于合一约束的汉语形式文法 · 句法结构歧义类型、歧义程度的定量考察、歧义消解策略	· 汉语短语结构关系系统 · 汉语词类、短语类体系 · 范畴原型：典型与非典型 · 范畴层级：大类与次类 · 中心语、补足语、附接语、指示语 · 语言成分组合中的功能变异	俞士汶等（1998/2003） 詹卫东（2000） 冯志伟（2010） R.K. Larson(2010) T.Walsh (2000) E. M. Bender (2013)
第 4 讲 语义知识表示及其应用	· 意义 = 形式变换 · 语义知识的类型：语义特征集、语义分类树、语义关系网 · 论元结构（配价关系） · 框架语义 · 意义组合性原则与情境性原则	· 义素分析 · 语义本体（ontology） · 论旨角色及其句法实现 · WordNet 、 FrameNet 、 HowNet 、 Propbank · AMR、UMR	冯志伟等译（2005）第 14、16 章 S.Lappin (1996)
第 5 讲 语篇知识表示及其应用	· 篇章连贯、篇章结构、指代 · 语段中心成分理论（CT） · 话语表示理论（DRT） · 修辞结构理论（RST）	· 语篇衔接的手段 · 约束指代成分的句法语义条件 · 语段潜在中心、现实中心、优选中心、语段跳转类型及其连贯性等级	Grosz, Joshi, and Weinstein(1995) 冯志伟等译（2005）第 18、19 章
第 6 讲 语料库的构建与应用	· 语料库的类型 · 语料库发展简史 · 语料库的设计与构建 · 语料库的应用	· 语料编码方案（XML、CES） · 语料标注工具（双语句子对齐） · 语料检索工具（concordance） · 树库（treebank）	黄昌宁、李涓子（2002） Gale & Church (1993) Marcus et.al.(1993)
第 7 讲 中文分词方法	· 由字到词 vs 由句到词 · 中文分词的困难 · 中文分词的基本方法 · 分词质量的评价	· **条件概率、贝叶斯公式、N-gram** · **动态规划算法**、编辑距离算法 · 词的内涵定义与外延定义（词表定义与语料库定义） · 交集型分词歧义、组合型分词歧义、	冯志伟等译（2005）第 5.6、6.2 节。 刘开瑛（2000） 黄昌宁、赵海（2007） 赵海等（2018）

		链长 · 未登录词的类型 · 最大匹配法、最大概率法、字位标注法分词 · 准确率、召回率、F-Score	
第 8 讲 词性标注方法	· 词的兼类现象 · 隐马尔可夫模型（HMM） · 维特比算法（Viterbi）	· 基于 HMM 的词性标注算法 · 基于转换的错误驱动词性标注算法	冯志伟等译（2005）第 8 章
第 9 讲 句法分析方法	· 句法分析技术的分类 · 自底向上分析：CYK 算法 · 自顶向下分析：Earley 算法 · 基于规则预处理的分析算法 Tomita 算法（GLR 算法） · 融合合一运算的句法分析	· Earley 算法的预测、扫描、归约算子及算法流程 · 合一运算与 Earley 算法的结合	冯志伟等译（2005）第 10、章、11.4、11.5 节、第 15 章
第 10 讲 机器翻译技术	· 机器翻译的分类 · 基于规则的机器翻译 · 基于实例（记忆）的机器翻译 · 基于统计的机器翻译 · 基于神经网络的机器翻译 · 机器翻译评测	· 机器翻译的需求类型 · 机器翻译中的人机关系类型 · 机器翻译的受控策略类型 · 机器翻译的实现技术类型 · 基于信道模型的机器翻译 · BLEU 机器翻译评测算法	冯志伟等译（2005）第 21 章

阅读文献

冯志伟（2010）《自然语言处理的形式模型》中国科学技术大学出版社。

冯志伟、孙乐译（2005）《自然语言处理综论》电子工业出版社（D.Jurafsky & J.H.Martin，*Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st Edition, 2000, Prentice-Hall Inc.）

黄昌宁、李涓子（2002）《语料库语言学》，商务印书馆。

黄昌宁、赵海，2007，中文分词十年回顾，《中文信息学报》2007 年第 3 期，8-19 页。

李维、郭进，2020，《自然语言处理答问》，商务印书馆，2020 年版。

刘开瑛，2000，《中文文本自动分词和标注》，商务印书馆。

俞士汶 等（1998/2003）《现代汉语语法信息词典详解》（第二版），清华大学出版社、广西科学技术出版社 1998 年版。

詹卫东（2000）《面向中文信息处理的现代汉语短语结构规则研究》，清华大学出版社、广西科学技术出版社 2000 年版。

赵海、蔡登、黄昌宁、揭春雨，2018，中文分词十年又回顾：2007-2017（可从网上下载）

Asher, N. and Lascarides, N., 2003, *Logics of Conversation*. Cambridge University Press. （2010 年北京大学出版社影印出版）

Bender, Emily M., 2013, *Linguistic Fundamentals for Natural Language Processing*, Morgan & Claypool Publishers.

Gale, W. & Church,K., 1993, **A program for aligning sentence in bilingual corpora**, *Computational linguistics*, Vol.19, No.1.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein, 1995, **Centering : A Framework for Modelling the Local Coherence of Discourse**, *Computational Linguistics* 21(2),pp.203-225.

Lappin, Shalom, ed., 1996, *The Handbook of Contemporary Semantic Theory*, Oxford: Blackwell.

Larson, Richard K., 2010, *Grammar As Science*, MIT Press, 2010, Unit 23, pp.343-354.

Mann, William C. & Sandra A.Thompson, 1988, **Rhetorical Structure Theory: Toward a functional theory of text organization**, *Text*, Vol.8, No.3, pp.243-281.

Marcus, Mitchell P., Beatrice Santoriniy, Mary Ann Marcinkiewicz, 1993, **Building a large annotated corpus of English: The Penn Treebank**, *Computational Linguistics*, Vol.19, No.2.

Walsh, Thomas, 2000，*A Short Introduction to X-bar syntax and transformations*, 2nd edition, Parlay Press