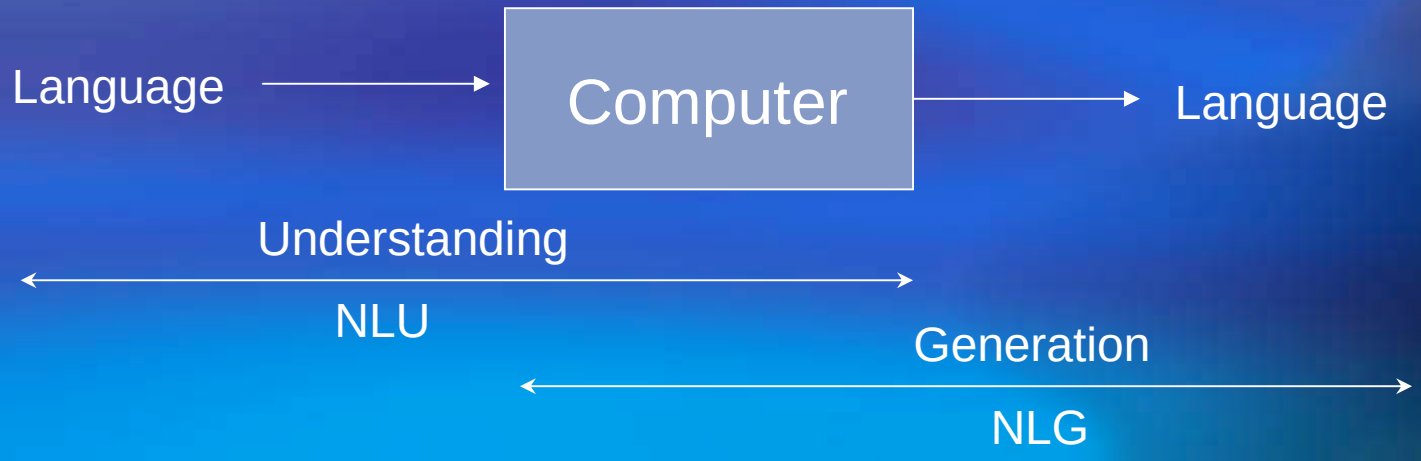

自然语言理解之——
中文信息处理
讲者：孙清

主要内容

- 自然语言理解介绍
- 国际上NLP的研究状况
- 中文信息处理的发展状况
- NLP的研究内容
- 今后的研究思路

自然语言处理的发展状况

- **Natural language processing** (**NLP**) is a subfield of artificial intelligence and computational linguistics. It studies the problems of automated generation and understanding of natural human languages.



Natural Language Processing

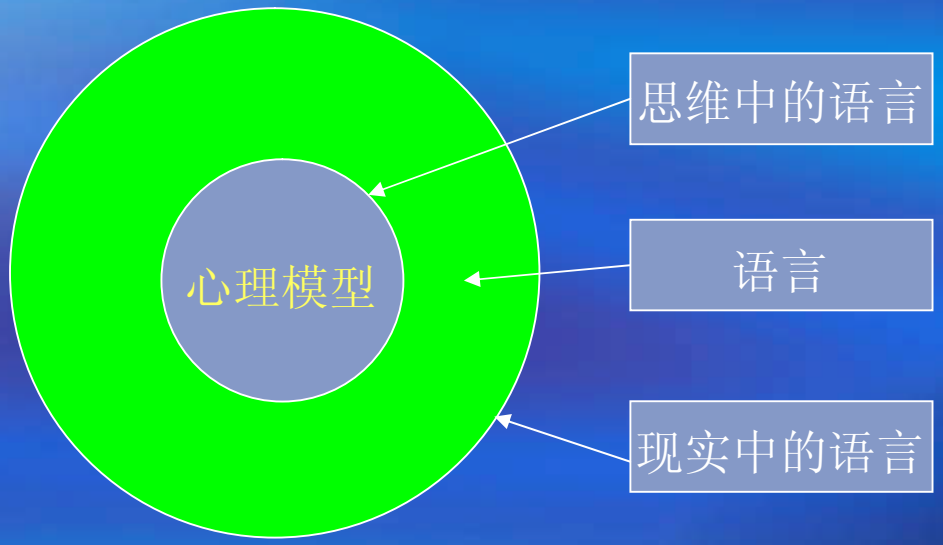
CL Discipline

Computational linguistics (CL) is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty.

It belongs to the cognitive sciences and overlaps with the field of ***artificial intelligence (AI)***

- Linguistics: the scientific study of language
- Computational: tool, modeling, and application
- Natural Language Processing
- Language Technology

语言在人脑中的形成



自然语言处理的不同层次

- 语音层(Phonetic Level)：研究词和其语音是如何相关联的，是语音处理的基础。
- 词法层(Morphological Level)：研究词是若何由意义的基本单位——词素构成的。
- 句法层(Syntactic Level)：研究词是如何组合成正确的句子的，词在句子中语法作用，以及哪些短语是其他短语的组成部分。
- 语义层(Semantic Level)：研究如何从一个句子中的词的意义，以及这些词在该句的语法结构中的作用来推导出该句的句义。语义分析是计算机理解自然语言的基础。
- 语用层(Pragmatic Level)：研究在不同的上下文环境中句子的使用。
- 话语层(Discourse Level)：研究前句对当前词义或句义的影响。

自然语言处理的主流理论介绍

- 扩充转移网络ATN： 基于图论数学概念的应用和语法研究的有限状态机。适用于语法分析。
- “格”语法：在深层结构中借用传统语法“格”的概念，来表示名词与谓语动词间一种固定不变的语义结构关系。适用于语法、语义分析。
- 概念依存理论CD：与格语法相似，句子意义表达以行为为中心，但句子的行为不由动词表示，而由原语行为集表示。适用于语法、语义分析和推理。
- 语义网络Semantic Network：依托深层结构理论，用结点表示词和短语的概念，用弧表示语义关系。适用于语义分析。（概念图、知识图）

Standardization in NLP

- As of October 1st, 2007 Common Logic is now a completed and published international standard referred to as "ISO/IEC 24707:2007 - Information technology — Common Logic (CL): a framework for a family of logic-based languages".

Common Logic

The ISO standard for Common Logic includes specifications for three dialects:

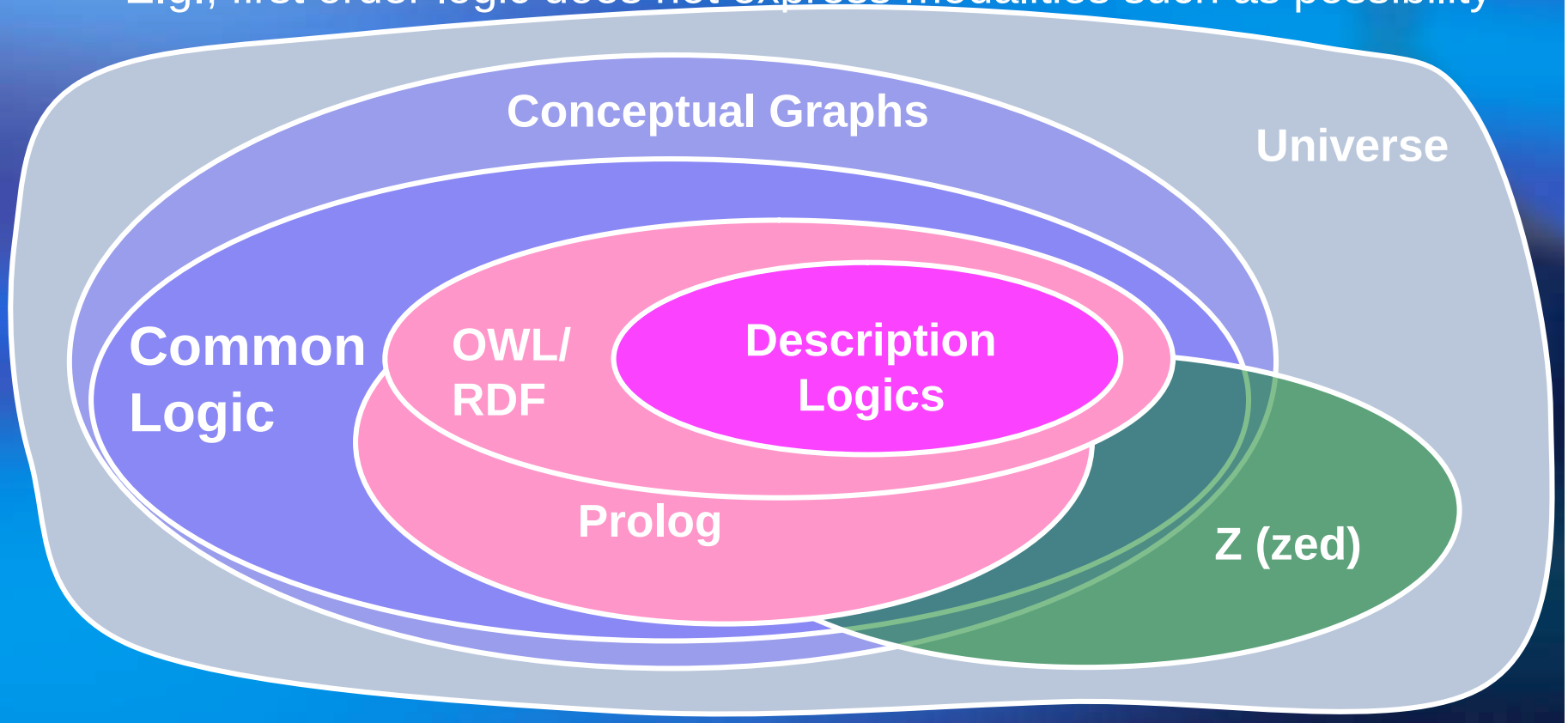
- the **Common Logic Interchange Format (CLIF)**
- the **Conceptual Graph Interchange Format (CGIF)**
- an XML-based notation for Common Logic **(XCL)**

The semantics of these dialects are defined by their translation to the abstract syntax and semantics of Common Logic.

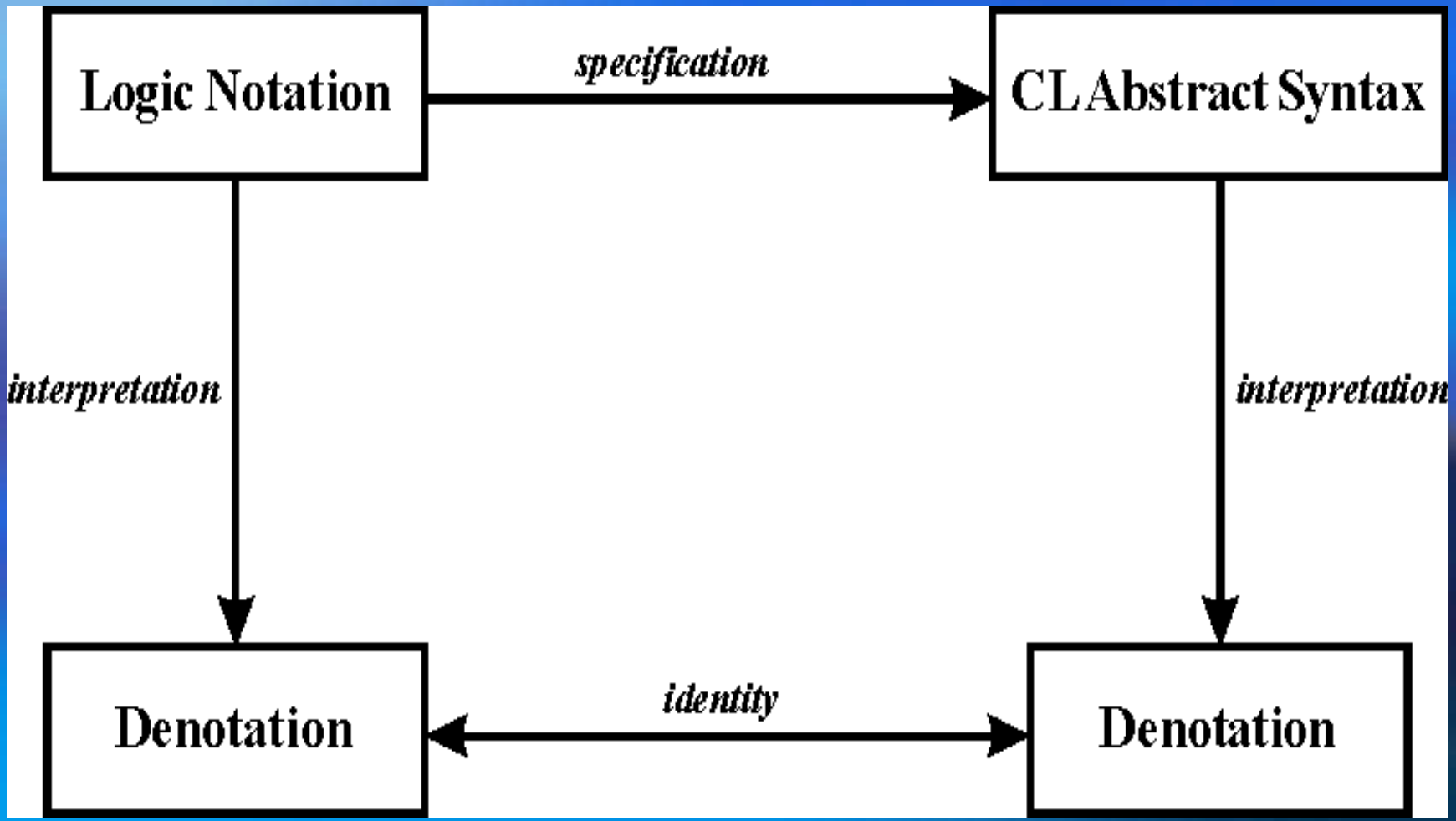
Many other logic-based languages could also be defined as subsets of CL by means of similar translations; among them are the RDF and OWL languages, which have been defined by the W3C.

Comparing Formalisms

- **Formalisms can be arranged by their expressivity ("power")**
  - The set of things that can possibly be expressed by the language
  - E.g., first order logic does not express modalities such as possibility



CL Conformance



Facilitate interoperability among logic-based languages used in several areas of IT:

- Metadata and ontology:
  - Knowledge Interchange Format (KIF).
  - Conceptual Graph Interchange Format (CGIF).
  - Cyc Knowledge Representation Language (CycL).
  - Description Logics (DLs).
- Semantic web:
  - Resource Definition Facility (RDF).
  - Web Ontology Language (OWL).
- Specification languages:
  - Unified Modeling Language (UML).
  - Z Formal Specification Notation (Z).

Software tools

- General Architecture for Text Engineering
- Natural Language Toolkit (NLTK): a Python library suite
- Expert System S.p.A.
- OpenNLP
- Bitext - The Bits and Text Company

中文信息处理的现状

- 现状和设想——试论中文信息处理与现代汉语研究（许嘉璐 2000年 ）
- 中文信息处理基本上还停留在“字处理”阶段
- 中文信息处理的三种思路：

第一个流派是以传统计算语言学为基本理论，从词素分析入手，进而研究词-短语（词组）-语段-句子。

第二个流派是HNC理论提出计算机对汉语的处理不应该以图灵检验为标准，而应该以对语言模糊的消解能力为第一标准

第三个流派是基于内涵模型论的语义分析

国内主要研究机构

- 清华大学智能技术与系统国家实验室
  自然语言处理组
- 北京大学 计算语言研究所
- 中国科学院 计算机与语言信息研究中心（知网）
- 中科院声学所 知识创新基地语言语音及交互信息技术部 （HNC）

中文语料库

- 人民日报切分/标注语料库 北京大学计算语言学研究所
- Sinica Corpus 现代汉语平衡预料库（台湾）
- The Lancaster Corpus of Mandarin Chinese (LCMC)
- 中文语义词库 CWB

  它含有 10 万以上的词条, 每个词条通过关系比较密切的相关词 (例如同义词、反义词、上位词、下位词等) 与其它词条相连结。整个词库呈现为比较复杂的网络结构, 并带有多种检索手段和显示方式。
- 知网（HowNet）——中文信息结构库

  包含268种信息结构模式，附带着一万多实例，总字数六万余。

语言信息处理的基础研究

- 面向信息处理应用的语言研究

  关于语言资源建设的专题研究

  语料的标注

  语料的分析和处理

  用于语言资源建设的字表、词表和标准、规范。
- 研究中的理性主义和经验主义方法

应用性的研究和实用系统的研制

- 机器翻译
  - 基于规则的方法和基于语料库的方法
  - 机器翻译中的专题研究
  - 应用型机器翻译系统的研制
- 文本信息处理

语言资源的建设

- 语料库
- 语言知识库
- 基于语料库的语言分析方法

今后的研究思路

- 学习基本的语言处理、知识表示的方法
- 掌握目前中文信息处理发展的状况
- 关注西文信息处理方面的最新发展
- 有针对性地对现有理论系统提出改进
- 面向典型应用构造实用系统的理论框架
- 开发实验平台。


谢谢！

Major tasks in NLP

- Automatic summarization
- Foreign Language Reading Aid
- Foreign Language Writing Aid
- Information extraction
- Information retrieval
- Machine translation
- Named entity recognition
- Natural language generation
- Optical Character Recognition
- Question answering
- Speech recognition
- Spoken dialogue system
- Text simplification
- Text to speech
- Text-proofing