Факультет компьютерных наук

Департамент анализа данных и искусственного интеллекта

Москва
2024

Лингвистические ресурсы

Лингвистический информационный ресурс

Корпусная лингвистика – область лингвистики, связанная с созданием и совершенствованием корпусов текстов, а также с их применением в качестве инструмента лингвистического исследования.

Корпус в лингвистике – это совокупность текстов, собранная в единое целое по определённым, соответствующим конкретной исследовательской задаче, критериям и отражающая ту или иную сферу использования языка.

Множество определенным образом оформленных и организованных речевых и языковых данных, находящихся на машинных носителях и используемых в различных сферах деятельности человека.

Формирование лингвистических корпусов

Свойства:

- репрезентативность.
- пропорциональность.

Аннотирование, или разметка, – это описание каждой единицы текста с помощью специальной системы отметок. Основная цель аннотирования – снятие омонимии (грамматической, лексической и т.д.), т.е. добавление дополнительной информации о том, в каком значении употреблена та или иная форма. Например, в зависимости от контекста словоформе мою приписывается либо значение глагольной формы (мою – 1 л. ед.ч. н.вр.), либо формы притяжательного прилагательного (мою – ж.р. ед.ч. В.п.).

Небольшая часть разметки корпуса производится **вручную**, т.е. специалист описывает каждое слово в определённом фрагменте текста. Где возможно, процесс аннотирования **автоматизируется** с помощью специальных программ – морфологических и синтаксических анализаторов.

После разметки каждая единица текста корпуса вносится в базу данных, а к готовой базе подключается корпусный менеджер. **Корпусный менеджер** – это программа, предоставляющая удобный доступ к базе данных корпуса.

Каждый фрагмент текста (**контекст**) в выдаче корпусного менеджера, как правило, снабжается названием источника и гиперссылкой на весь текст, откуда он был взят. Корпусный менеджер может сообщать разного рода **статистическую информацию** по корпусу: например, строить частотный словарь употребления той или иной единицы текста.

Основные составляющие

Первичные – устная и письменная речь без изменений (например, первое издание «Руслана и Людмилы»).

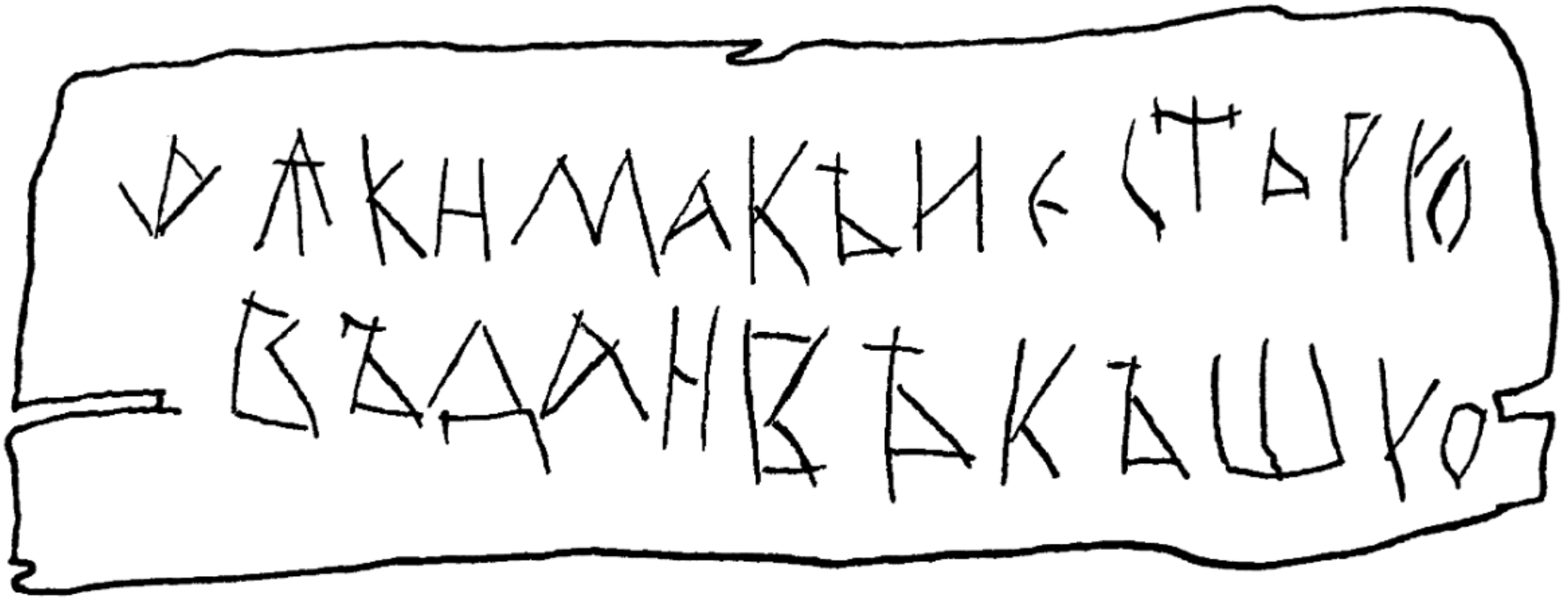
Вторичные – результат оценивающей деятельности на основе первичных данных.

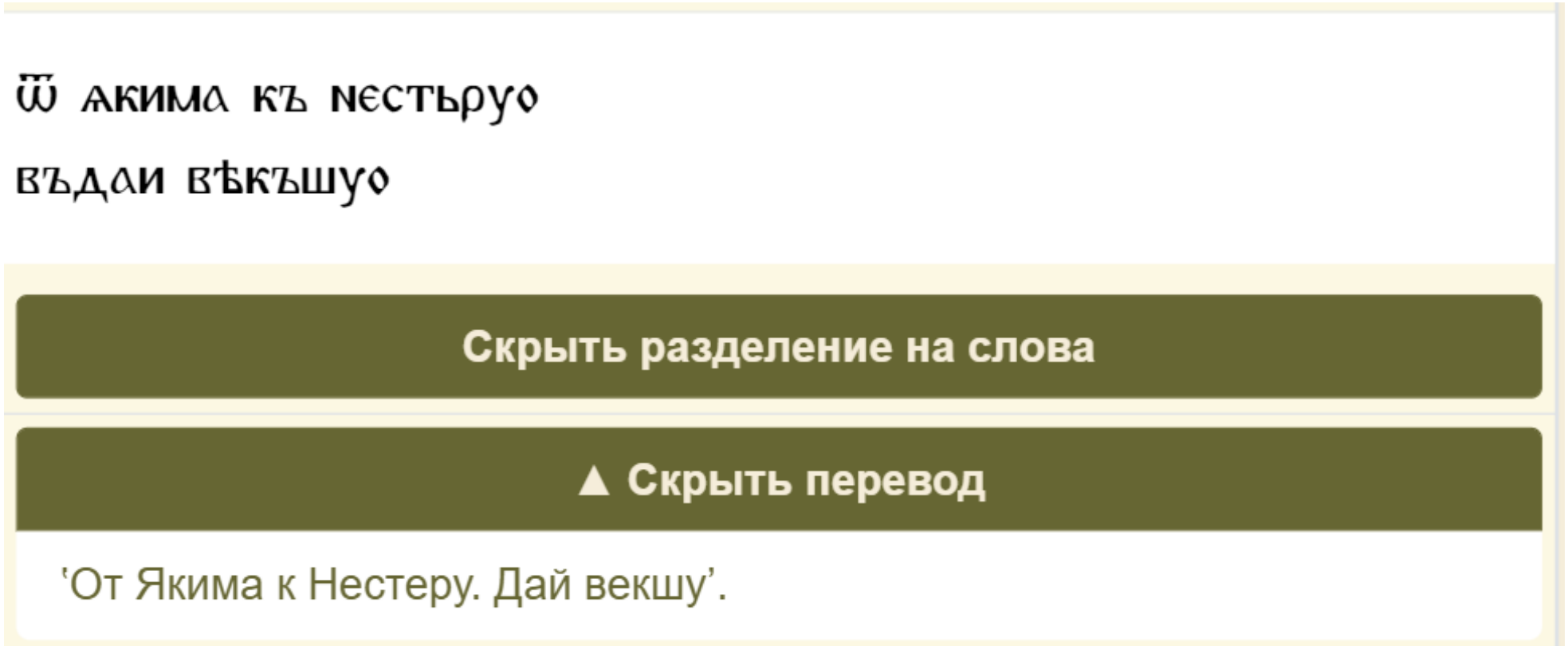
- Текстовые корпуса
- Словари (одноязычные и многоязычные)
- Терминологические словари и БД
- Ресурсы устной речи

Виды лингвистической разметки

- **Морфологическая –** части речи.
- **Синтаксическая –** грамматика.
- **Семантическая –** значения.
- **Анафорическая –** связи.
- **Просодическая –** ударения и интонации.
- **Структурная –** сегменты текста.

I (pronoun) will (verb) use (verb) Google (noun) before (preposition) asking (verb) dumb (adjective) questions (noun) .



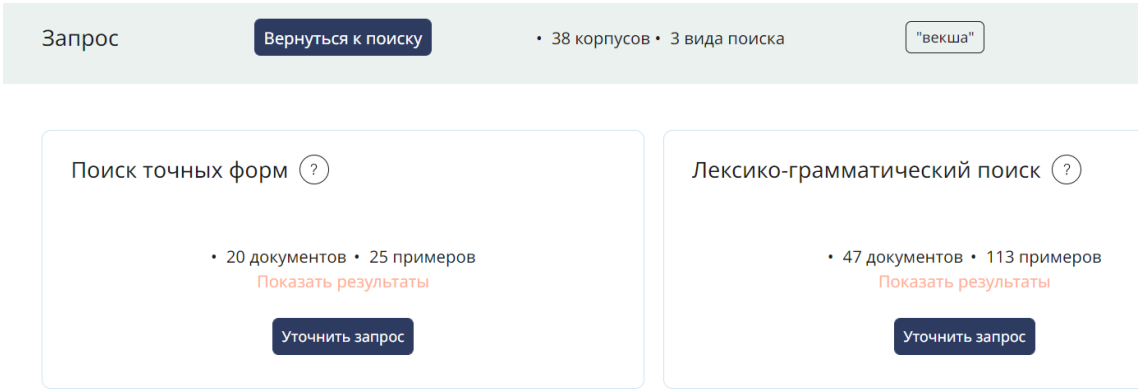

ѿ акима къ нєстьруо
въдаи вѣкъшуо
Скрыть разделение на слова
▲ Скрыть перевод
‘От Якима к Нестеру. Дай векшу’.

Национальный корпус русского языка (НКРЯ)

Собрание независимых корпусов, каждый из которых предназначен для решения определенных лингвистических задач. Каждая из этих коллекций текстов является большой по объёму и представительной, что делает их ценным материалом для количественных и качественных исследований.

Помимо текстов на современном русском языке, ориентированном на литературный стандарт, НКРЯ стремится представить русский язык в его историческом и географическом многообразии.

Большинство корпусов, входящих в НКРЯ, одноязычные, то есть в них входят только тексты на одном языке. Исключением является параллельный корпус, где оригинальные русские тексты сопровождаются переводом на другой язык, или иноязычные произведения переведены на русский.



WordNet & FrameNet

WordNet

- A semantic lexicon for the English language
- Purpose:
  - A combination of dictionary and thesaurus
  - to support automatic text analysis and artificial intelligence applications

WordNet

- Groups the meanings of English words into five categories
  - Nouns
  - Verbs
  - Adjectives
  - Adverbs
  - Function words(prepositions, pronouns, determiners)

WordNet

- Meanings are related by
  - Synonymy (Pipe, Tube)
  - Antonymy (Wet, Dry)
  - Hyponymy (Tree, Plant)
  - Meronymy (Ship, Fleet)
  - Morphological relations

Application - WordNet

- WordNet’s hierarchical structure can help in the creation faceted categories, which are essential for faceted metadata and search functions.
- Words from a structured collection are compared to high-level category labels of WordNet’s lexicon.
  - Subsets of the most frequently occurring categories are retained.
  - Categories related to ambiguous words are discarded.
  - High-level hierarchy labels that are too general or broad are discarded as well.

Application - WordNet

- **Reason for using WordNet?**
  - Allows for efficient navigation within and across lexical data due the rigorous structure of its semantic tagging
  - Hypernym (IS A) relations are most commonly used and easiest to integrate into Information Extraction and browsing/search systems, making it easier to find synonyms and near synonyms of words.
  - Currently, there has been a movement to create multilingual WordNets with the goal of enhancing cross-lingual information retrieval systems. WN provides a platform for representing the lexical knowledge between different languages.

Data Structure & Maintenance

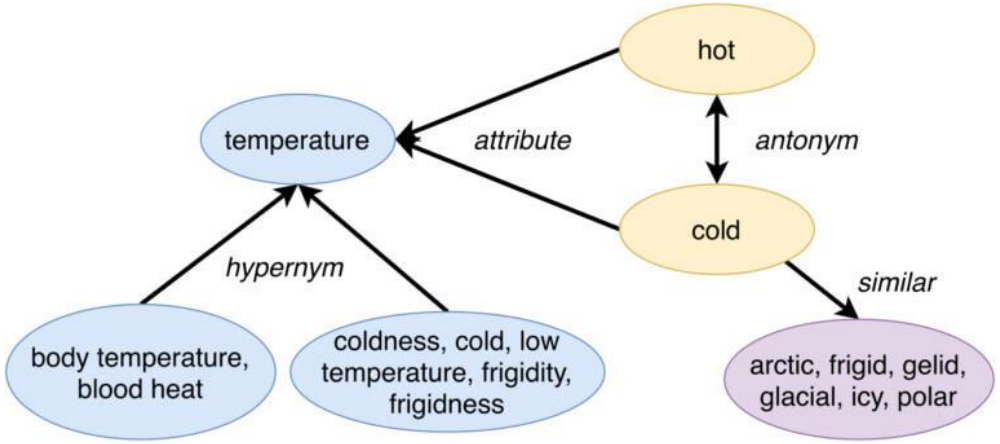
- WordNet was created and is being maintained at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George A. Miller
- Development began in 1985
- Q: Where do they get the definitions for WordNet?
- A: Their Lexicographers write them
  - However, many different dictionaries and sources were used and many others are still being used to expand the WordNet library.
- The database contains about 150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs

(Ru)WordNet

Тезаурус RuWordNet содержит синсеты (наборы синонимов) трех частей речи: **существительные** (отдельные существительные, группы существительного, предложные группы), **глаголы** (отдельные глаголы и глагольные группы), **прилагательные** (отдельные прилагательные и группы прилагательного).

Всего тезаурус RuWordNet содержит 111.5 тысяч слов и выражений русского языка.

Между синсетами, относящимися к разным частям речи, но выражающими один и тот же смысл, установлены отношения частеречной синонимии, соединяющие разделенные синсеты. Также между синсетами установлены отношения: гипоним-гипероним (род-вид), экземпляр-класс, отношение антонимии, часть-целое, причина, логическое следование, предметная область (домен).

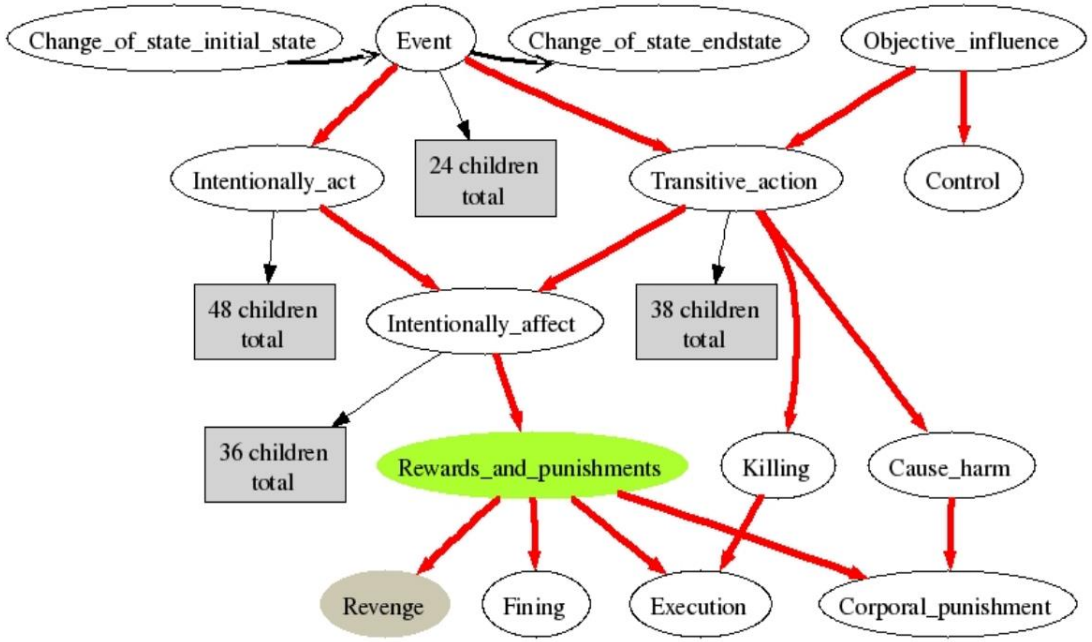


FrameNet

- A project housed at the International Computer Science Institute in Berkeley, California which produces an electronic resource based on semantic frames
- Scope of the project
  - FrameNet Database : Lexicon, Frame Database, Annotated Example Sentences
  - Associated Software Tools

FrameNet

Фрейм — это схематическое изображение ситуации с участием различных участников, реквизита и других концептуальных ролей. Примеры имен фреймов: «Бытие_рождения» и «Местное_отношение». Фрейм в FrameNet содержит текстовое описание того, что он представляет (определение фрейма), связанные элементы фрейма, лексические единицы, примеры предложений и отношения между фреймами.



FrameNet

frame(TRANSPORTATION) frame_elements(MOVER(S), MEANS, PATH) scene(MOVER(S) move along PATH by MEANS)
frame(DRIVING) inherit(TRANSPORTATION) frame_elements(DRIVER (=MOVER), VEHICLE (=MEANS), RIDER(S) (=MOVER(S)), CARGO (=MOVER(S))) scenes(DRIVER starts VEHICLE, DRIVER controls VEHICLE, DRIVER stops VEHICLE)
frame(RIDING_1) inherit(TRANSPORTATION) frame_elements(RIDER(S) (=MOVER(S)), VEHICLE (=MEANS)) scenes(RIDER enters VEHICLE, VEHICLE carries RIDER along PATH, RIDER leaves VEHICLE)

FrameNet

FEG	Annotated Example from BNC
D	[$_D$ Kate] **drove** [$_P$ home] in a stupor.
V, D	A pregnant woman lost her baby after she fainted as she waited for a bus and fell into the path of [$_V$ a lorry] **driven** [$_D$ by her uncle].
D, P	And that was why [$_D$ I] **drove** [$_P$ eastwards along Lake Geneva].
D, R, P	Now [$_D$ Van Cheele] was **driving** [$_R$ his guest] [$_P$ back to the station].
D, V, P	[$_D$ Cumming] had a fascination with most forms of transport, **driving** [$_V$ his Rolls] at high speed [$_P$ around the streets of London].
D+R, P	[$_D$ We] **drive** [$_P$ home along miles of empty freeway].
V, P	Over the next 4 days, [$_V$ the Rolls Royces] will **drive** [$_P$ down to Plymouth], following the route of the railway.

Figure 2: Examples of Frame Element Groups

FrameNet

- Comparison with WordNet and Ontology
  - Lexical units comes with definition
  - Multiple annotated example
  - Examples from natural corpora
  - Frame by frame
  - A network relations between frames
  - Not readily usable as ontology of things

Application - FrameNet

- Organize information in terms of case-roles, which helps determine the lexical meaning by the use of conceptual structure provided by FN.
- Can be applied to NLP systems because of its potential to find the arguments of a collection through the use of word sense and sentence examples.
- FrameNet annotated data sets are compared against Information extraction patterns. All non-relevant terms of the frames are discarded.

Application - FrameNet

- **Reason for using FrameNet?**
  - The lexicon and pattern sets provided by FN make it possible for natural language processing systems to generate more precise results than those allowed by WordNet.
  - FN consists of machine readable terms that provide sentence examples extracted from natural corpora, which make it possible to provide meaning to terms related to frames.

Data Structure

- The development of the theory of Frame Semantics began more than 25 years ago, however until 1997 there were no implementations
- British National Corpus and Linguistic Data Consortium were used to create the database and they plan to add American National Corpus data as well
- Frames are added by FrameNet Staff

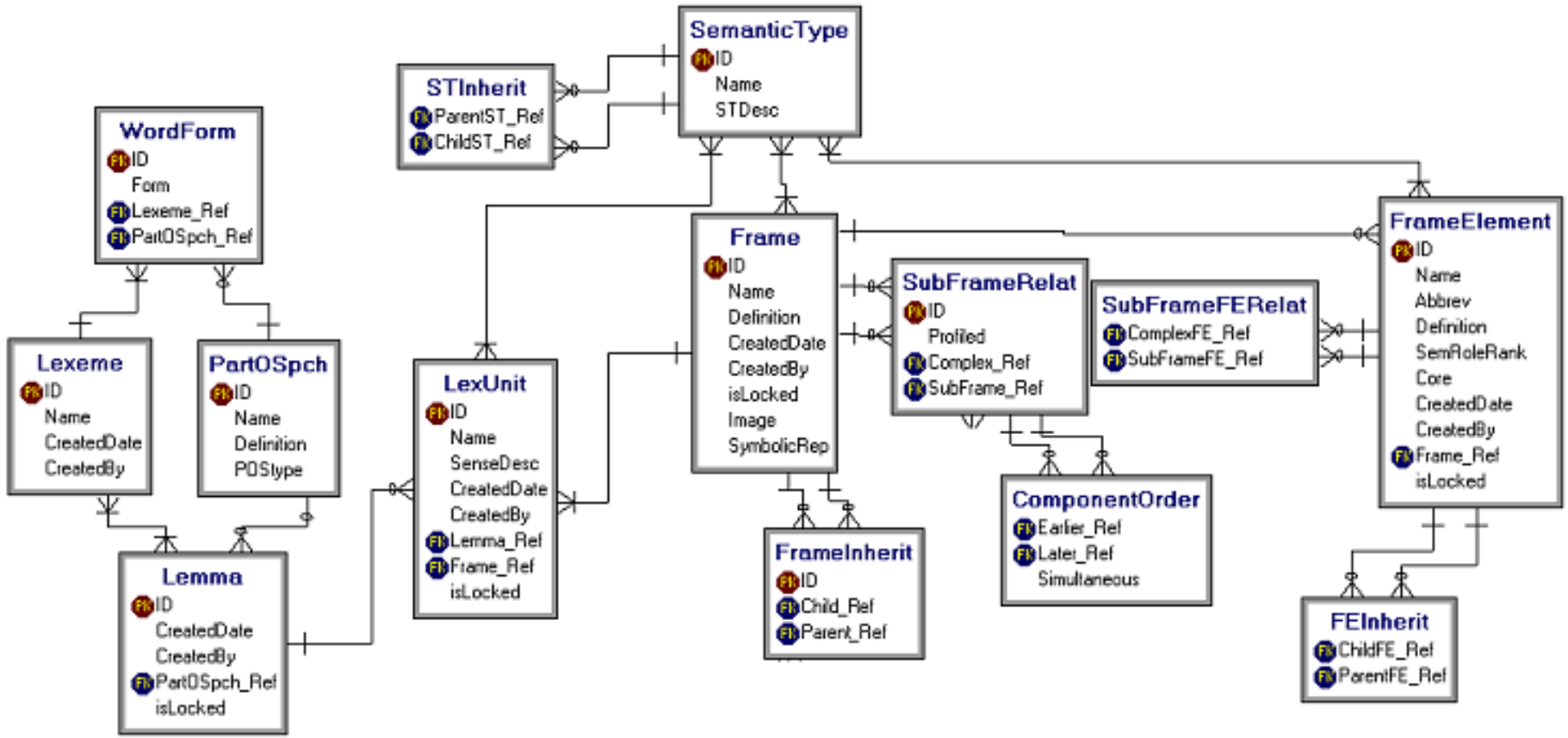


Figure 1: Structure of the frame database

Authority

Lexical Unit

Definition:

An Agent has the means to affect a Theme along the lines of a certain Domain. There is an imbalance of influence or power within a certain Domain that favors the Agent (often due to some Source) over the Theme.

His AUTHORITY over his subjects was given to him by God.

Congress has the POWER to enact laws.

Only the King has the AUTHORITY to declare war.

FEs:

Core:

Agent [Agt]	The person or organization with more influence or power.
Theme [Thm]	The person or organization with less influence or power.

Relevance with IE

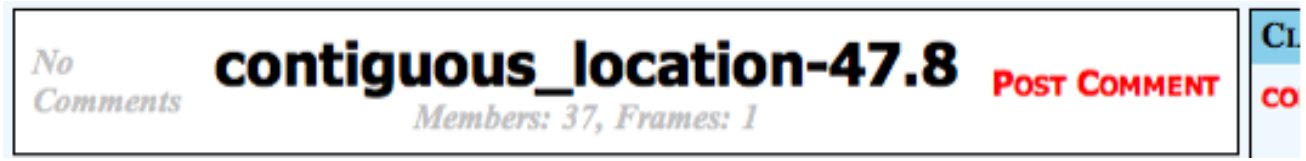
- FN and WN are essential resources for Natural Language Processing applications and Information Extraction systems.
- FN and WN have been used for information retrieval, word sense disambiguation, machine translation, conceptual indexing, and text and document classification, among other applications/systems.

VerbNet & PropBank

VerbNet

- Organizes verbs into classes that have common syntax/semantics linking behavior
- Classes include…
  - A list of member verbs (w/ WordNet senses)
  - A set of thematic roles (w/ selectional restr.s)
  - A set of frames, which define both syntax & semantics using thematic roles.
- Classes are organized hierarchically

VerbNet - *cover contiguous_location-47.8*



Full Class View

contiguous_location-47.8
- contiguous_location-47.8-1
- contiguous_location-47.8-2

Member Verb Lemmas:

BESTRIDE, BLANKET, BORDER, BOUND, BRACKET, BRIDGE, CAP, CIRCLE, CLOAK, COVER, EDGE, ENCIRCLE, ENCLOSE, ENCOMPASS, ENGULF, ENVELOP, FENCE, FILL, FLANK, FOLLOW, FORGO, FRAME, HEAD, HUG, LINE, NEIGHBOR, OVERCAST, OVERHANG, PRECEDE, PREDATE, RIM, RING, SKIRT, SPAN, STRADDLE, SUPPORT, SURMOUNT, SURROUND, TOP, TRAVERSE, UNDERLIE, WREATHE, ENSHROUD, VEIL

ROLES:

Theme [+concrete]
Co-Theme [+concrete]

NP V NP

EXAMPLE:

Italy borders France.

SHOW DEPENDENCY PARSE TREE

VerbNet Thematic Roles

- Actor
- Actor1
- Actor2
- Agent
- Asset
- Attribute
- Beneficiary
- Cause
- Destination
- Experiencer
- Extent
- Instrument
- Location
- Material
- Patient
- Patient1
- Patient2
- Predicate
- Product
- Proposition
- Recipient
- Source
- Stimulus
- Theme
- Theme1
- Theme2
- Time
- Topic
- Value

PropBank

- 1M words of WSJ annotated with predicate-argument structures for verbs.
  - The location & type of each verb's arguments
- Argument types are defined on a per-verb basis.
  - Consistent across uses of a single verb (sense)
- But the same tags are used (Arg0, Arg1, Arg2, …)
  - Arg0 ≈ proto-typical agent (Dowty)
  - Arg1 ≈ proto-typical patient

PropBank Example:
cover (smear, put over)

- Arguments:
  - Arg0 = causer of covering
  - Arg1 = thing covered
  - Arg2 = covered with
- Example:

  John ***covered*** the bread with peanut butter.

PropBank:
Trends in Argument Numbering

- Arg0 = proto-typical agent (*Dowty*)
  Agent (85%), Experiencer (7%), Theme (2%), …
- Arg1 = proto-typical patient
  Theme (47%),Topic (23%), Patient (11%), …
- Arg2 = Recipient (22%), Extent (15%), Predicate (14%), …
- Arg3 = Asset (33%), Theme2 (14%), Recipient (13%), …
- Arg4 = Location (89%), Beneficiary (5%), …
- Arg5 = Location (94%), Destination (6%)

(Percentages indicate how often argument instances were mapped to VerbNet roles in the PropBank corpus)

PropBank: Adjunct Tags

- Variety of ArgM's (Arg#>5):
  - TMP: when?
  - LOC: where at?
  - DIR: where to?
  - MNR: how?
  - PRP: why?
  - REC: himself, themselves, each other
  - PRD: this argument refers to or modifies another
  - ADV: others

Lexical Resources (short summary)

- **WordNet** -- What verbs are synonymous?
- **FrameNet** -- How do verbs that describe a common scenario relate?
- **VerbNet** -- How do verbs w/ shared semantic & syntactic features (and their arguments) relate?
- **PropBank** -- How does a verb relate to its arguments? Includes **annotated text.**