CL/NLP Basics

Lecture 2

Good! Feedbacks are already coming in!

- Better Recordings!
  - Thanks to Vinayak!
- Last row of seats reserved for DB victims.
  - For people who arrived late.
  - If you arrive late, please squeeze in the last row.
- Please do not leave or pack your items in the **last five minutes**.
- Zoom available.
  - However, we will have prioritize people who attend in person.
  - Can’t guarantee streaming quality & interaction.

Announcements

- HW1 released!
  - Due on Nov 6 (Thursday) 4pm.
- Two lectures this week:
  - Swapping this Thursday’s lecture with next Tuesday’s lecture.
  - Lectures: Today and Thursday (in the original rooms).
  - **23.10.2025: 16:15 - 17:55, Room: S103/221** → Lecture
  - **28.10.2025: 13:30 - 15:10, Room: S306/051** → Practice Session

Outline

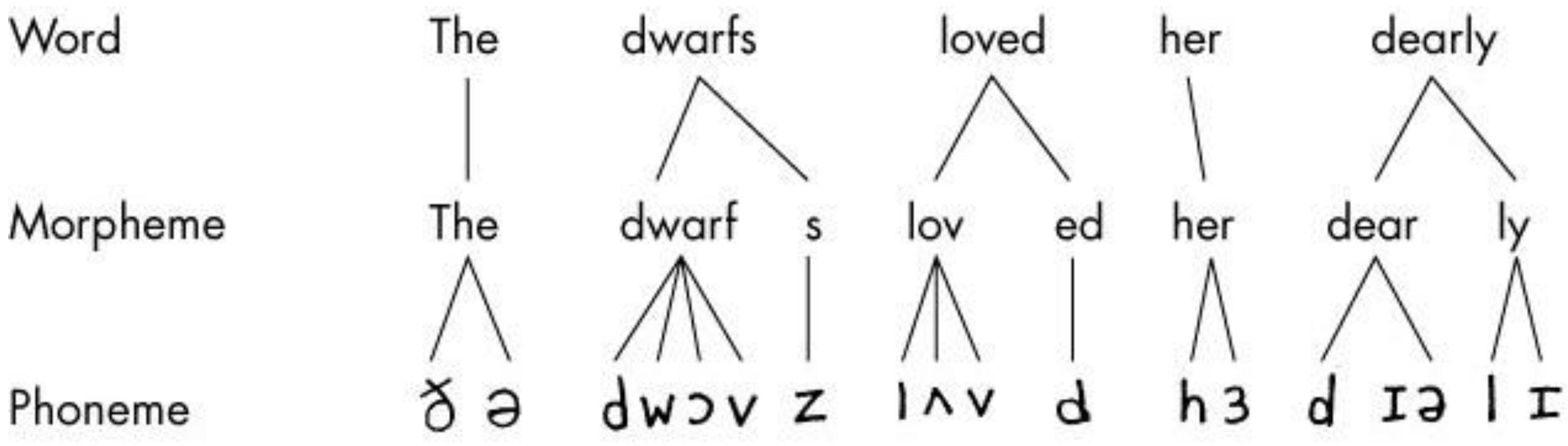
- Linguistics basics.
- Parsing:
  - Constituency parsing.
  - Dependency parsing.
- Word sense disambiguation (WSD).
  - WordNet
  - Lesk’s algorithm

Some Linguistics

- **Linguistics** is the scientific study of language.
- “Levels” of study:
    - Phonetics, phonology, morphology, syntax, semantics, pragmatics and discourse.
    - No agreement on how many total levels: 4, 5, 6, 7…
    - Also debate on whether we should use the idea of **“levels”** to describe language structure.
- “Area” of linguistics may be more appropriate.

Phonetics & Phonology

- Phonetics
  - How speech sounds are physically produced and perceived.
- Phonology
  - How those sounds are mentally organised and patterned within a language.



Morphology

- Morphology is the study of **words**.
- Morphology is the study of the minimal meaningful units of language.
- Inflectional morphology:
    - How words change forms.
    - E.g., walk, walks, walked.
- Derivational morphology:
    - How words become new words.
    - E.g., happy → unhappy; teach → teacher; hospital → hospitalised.
    - Darmstadt → darmstadty → darmstadtify → darmstadtifisation → dedarmstadtifisation → post-darmstadtifisation → post-darmstadtifisationism.

Syntax

- The study of word orders.
- "Syntax is the study of the regularities and constraints of word order and phrase structure“
  (Manning & Schütze, 2003, p. 93)
- The combinatorial structure of words.
- How words can be linearly organized:
  - ***Left/right precedence*.**
  - ***Contiguity***.
- How words can be hierarchically organized into ***phrases*** and ***sentences***.

Syntactic structure

- *Syntax*:

The cat hunted the squirrel living in the tree with persistence.

vs.

Squirrel persistence in cat the living tree with the hunted the.

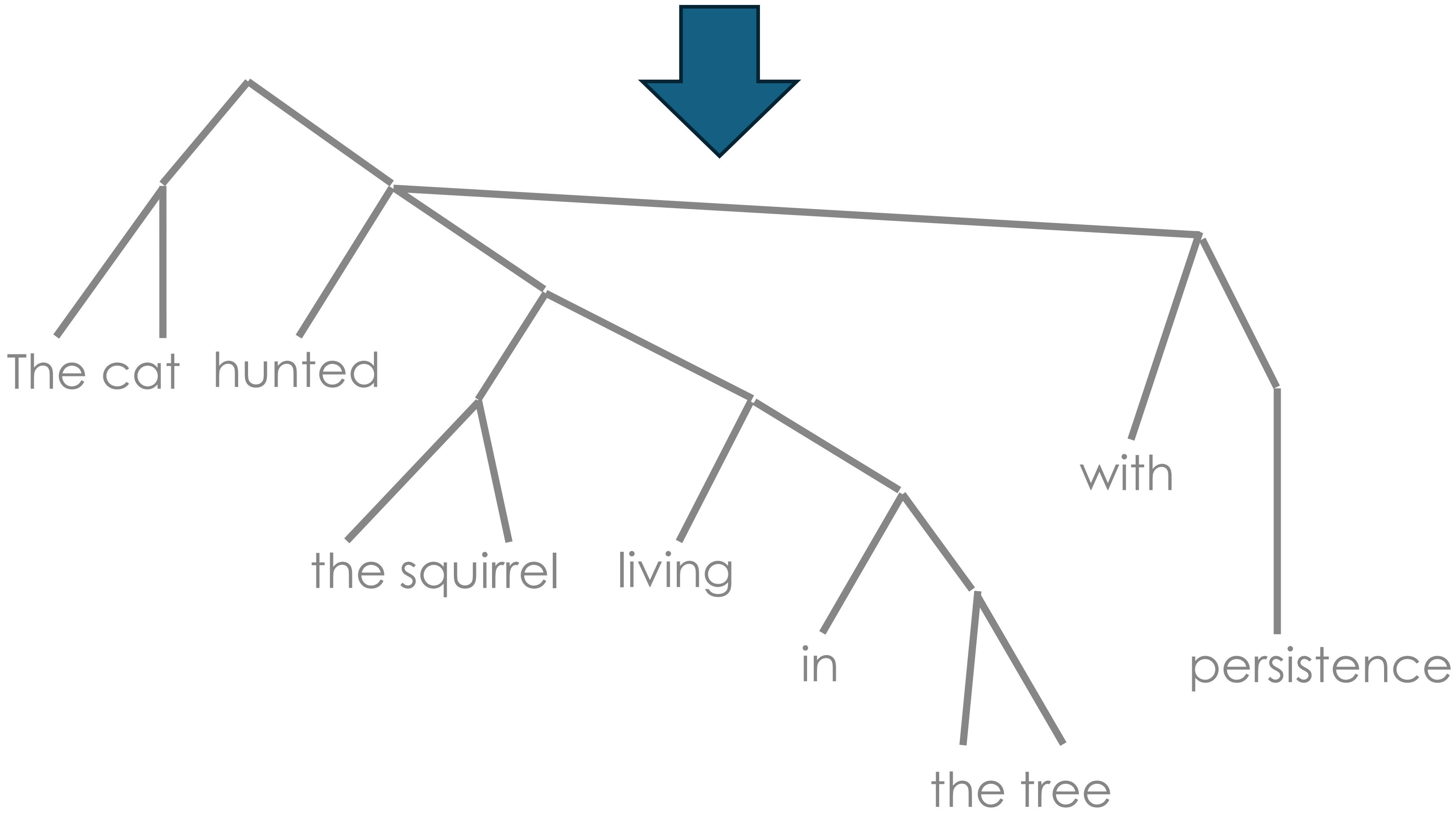
Syntactic structure

The cat hunted the squirrel living in the tree with persistence.

```
[[The cat]
 [hunted [the squirrel [living [in [the tree]]]]
 [with [persistence]]]]
```

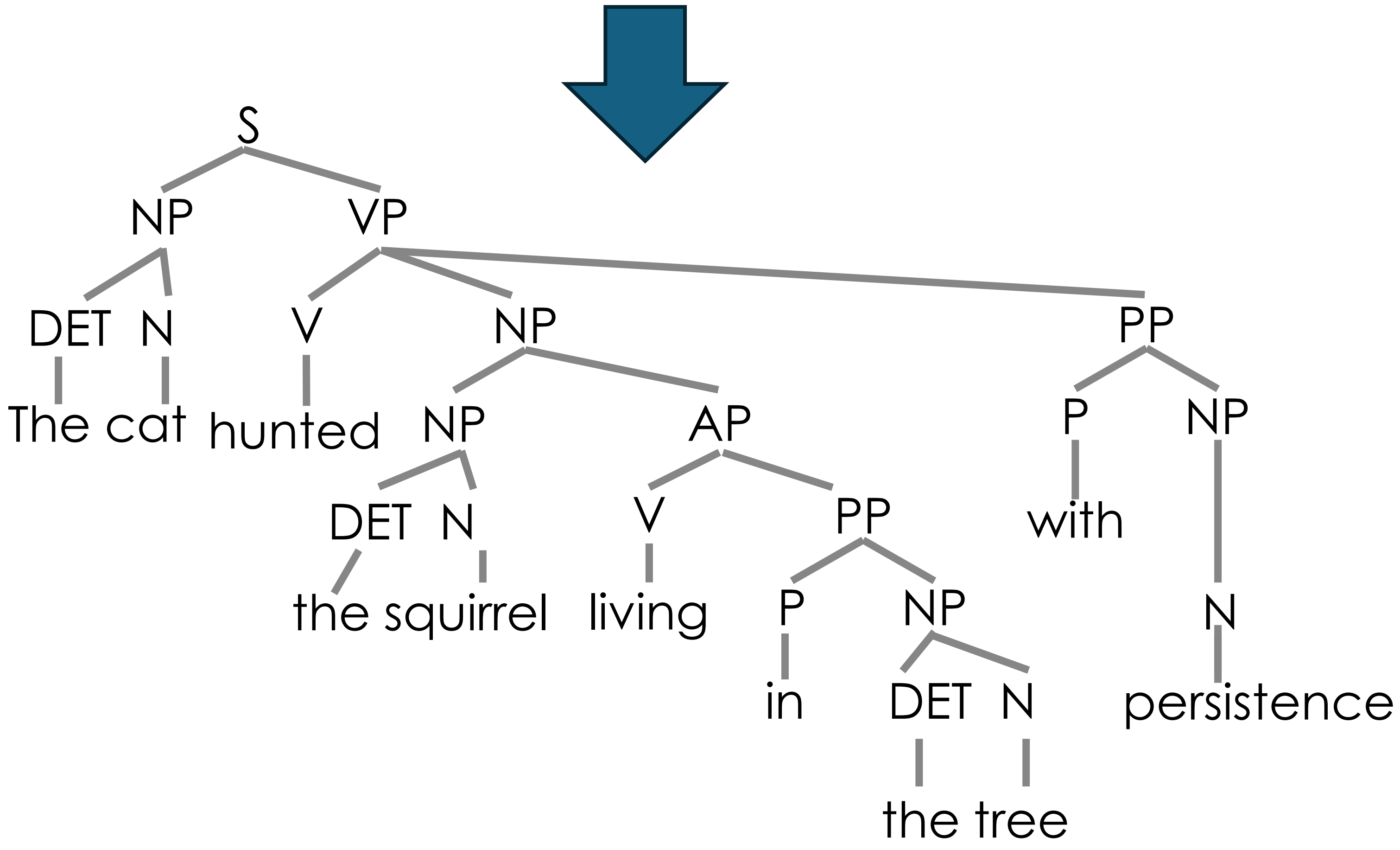
Syntactic structure

The cat hunted the squirrel living in the tree with persistence.



Syntactic structure

The cat hunted the squirrel living in the tree with persistence.



Grammars and Parsing

- Grammar:
    - Formal specification of allowable structures.
        - Knowledge
        - Representation
- Parsing:
    - Analysis of string of words to determine the structure assigned by grammar.
        - Algorithm
        - Process

Semantics

- The study of meaning.
- Lexical semantics:
  - Word meanings and their internal structure.
  - And, the structure of the relations among words and meanings.

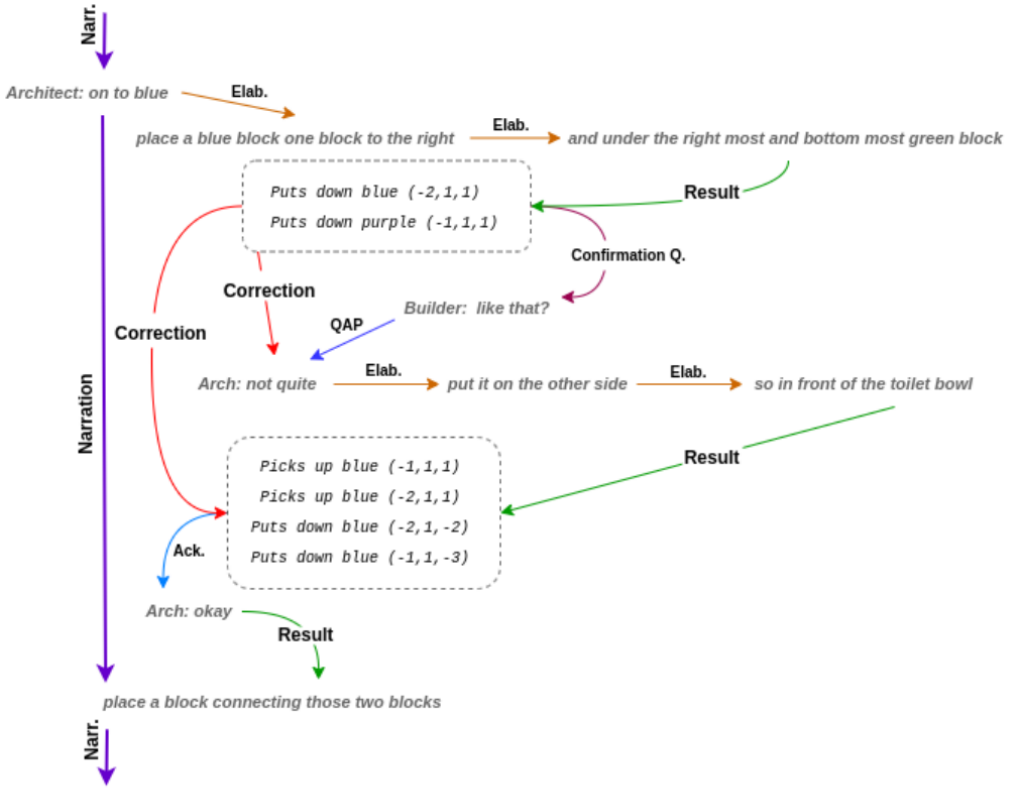
- Every student takes a course.
  - $\exists y.(\mathtt{course}(y) \wedge \forall x.(\mathtt{student}(x) \Rightarrow \mathtt{take}(x, y)))$
  - $\forall x.(\mathtt{student}(x) \Rightarrow \exists y.(\mathtt{course}(y) \wedge \mathtt{take}(x, y)))$

Pragmatics & Discourse

- Pragmatics: study of language in use.

- John: I find linguistics pretty boring.
- Mary: I find it boring, too.
- Mary: I find it boring.

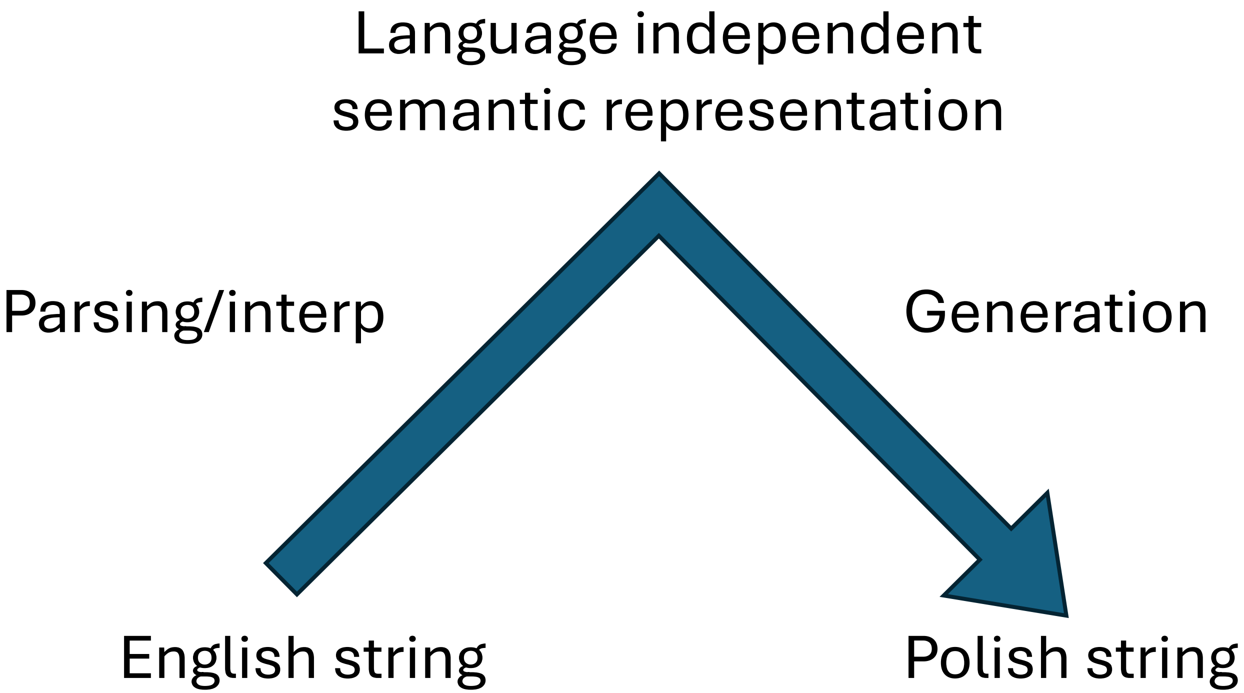
- Discourse: study of the structure of text and dialogues.



Thompson et al. (2024). Discourse Structure for the Minecraft Dialogue Corpus.

Wait, how is all these relevant?

- Machine Translation:
- the Vauquois triangle (1968).



Language independent semantic representation

Parsing/interp

Generation

English string

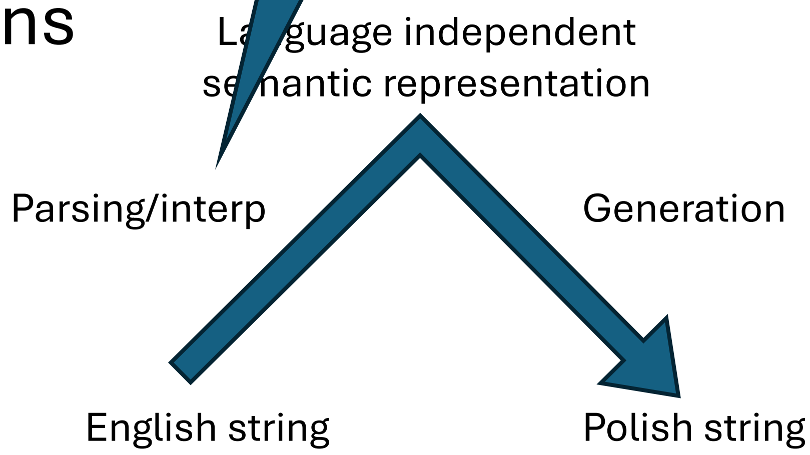
Polish string

Take MT as an example

- Tokenization for splitting texts into tokens
    - [‘Birds’, ‘does’, ‘not’, ‘exist’]
- PoS-Tagging and parsing analyse syntactic features
    - N, V, NegP, V
    - [S [NP Birds] [VP does [NegP not [VP exist]]]]
- Stemming / Lemmatization to normalize tokens

    -
      ```
      ( e / exist-01
            :polarity –
            :arg1 ( b / bird ) )
      ```
- Generation
    - Ptaki nie istnieją

At the center: grammar and parsing!



Tokenization

- Segmenting an input stream into an ordered sequence of units is called **tokenization**.
- A system which splits texts into tokens is called a **tokenizer**

A very simple example:

- Input text:
  John likes Mary and Mary likes John.
- Tokens:
  {"John", "likes", "Mary", "and", "Mary", "likes", "John", "."}

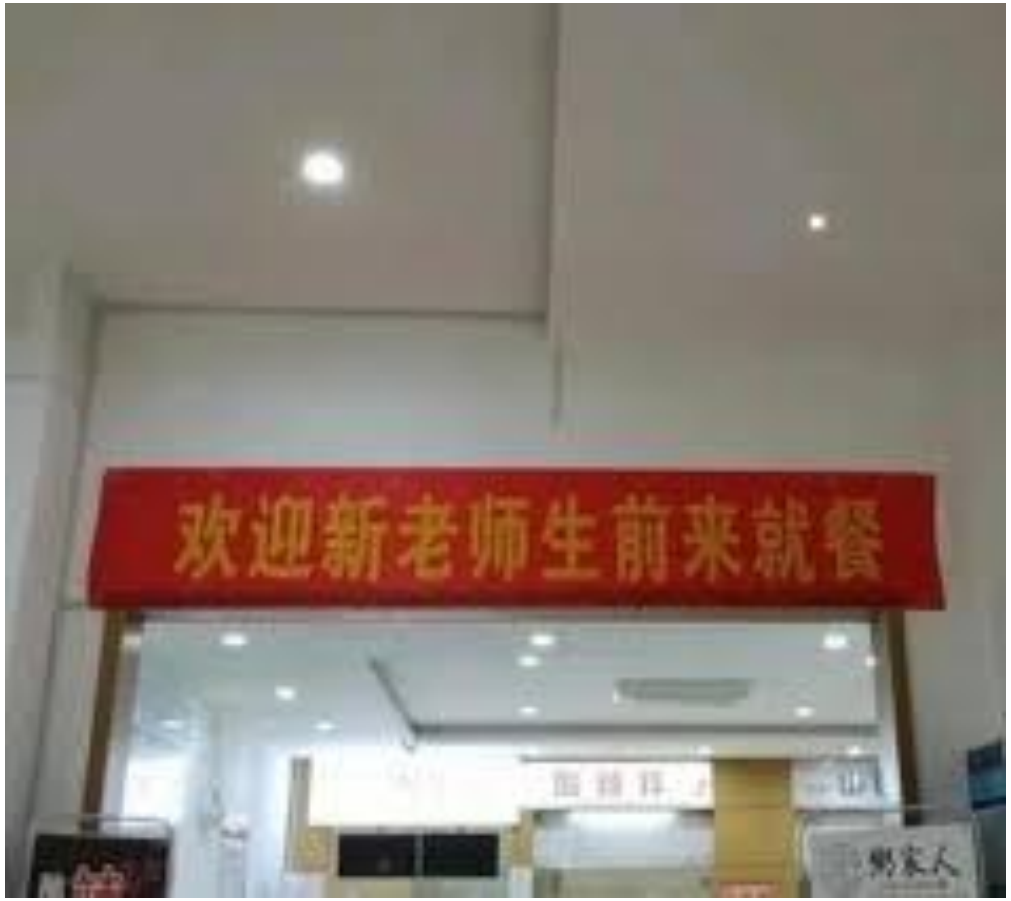
Surprisingly hard for English.

Tokenization

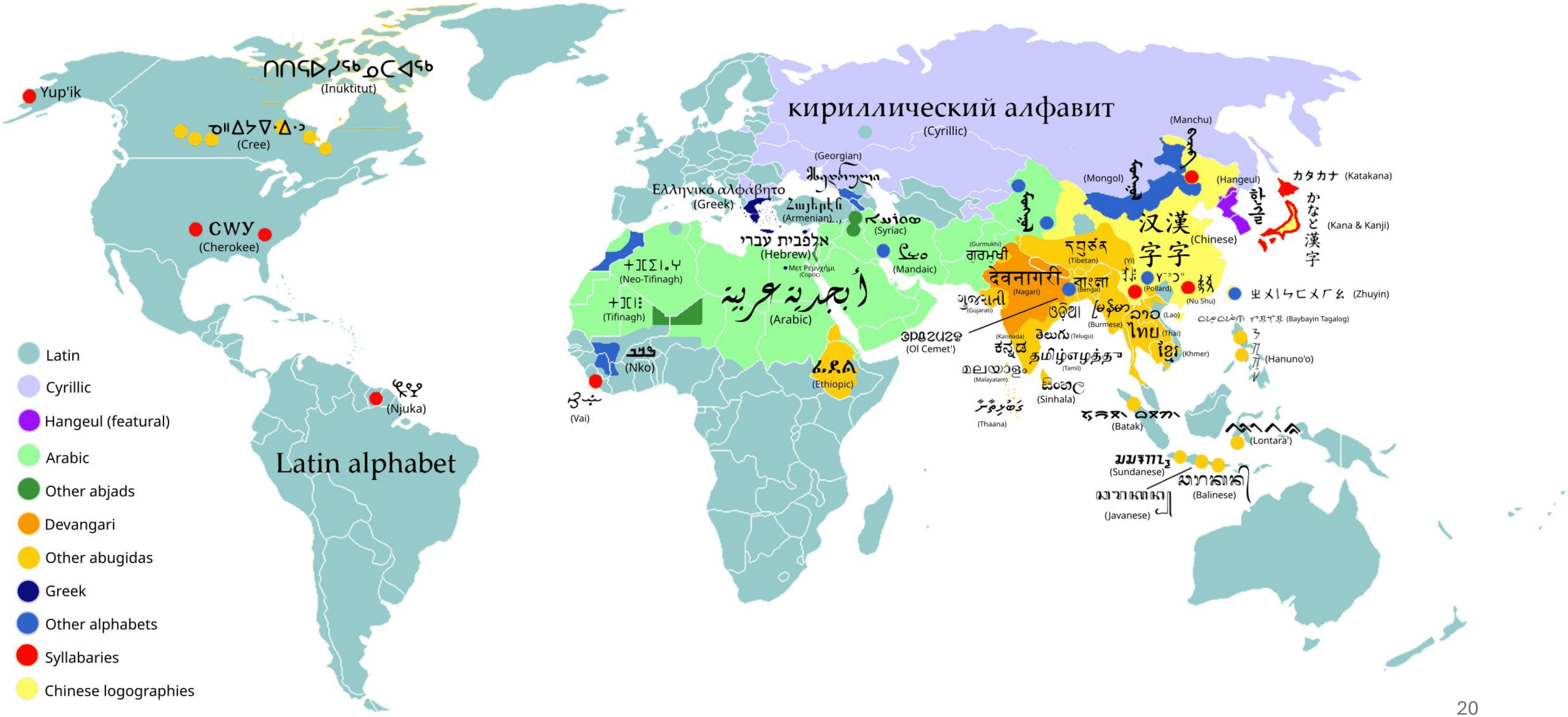
- Tokenization ambiguity in Chinese.
- 新老 / 师生 / 前来 / 就餐
- new old / teacher student / come forward / to dine
- New and old teachers and students come to dine.

- 新老师 / 生前 / 来就餐
- New teacher / before died / come dine
- The new teacher came to eat *before they died*

- Surprisingly easy in nowadays.



Writing Systems



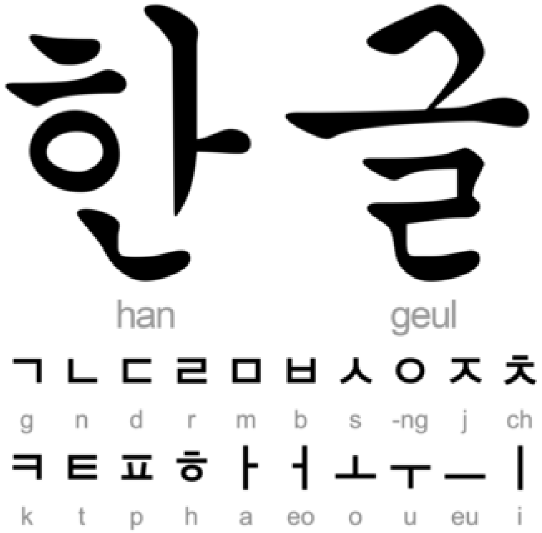
Major Writing Systems

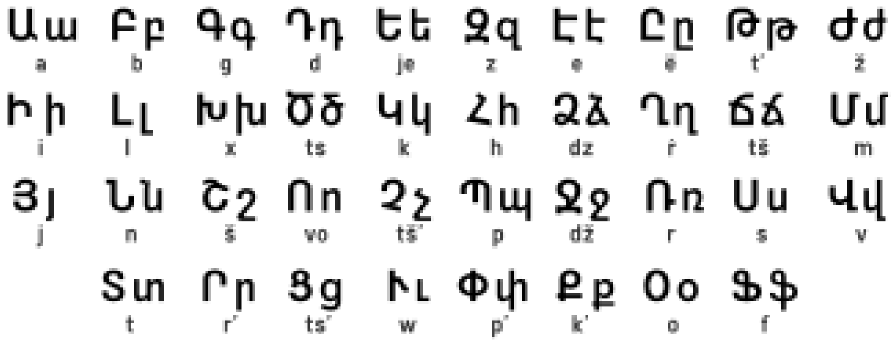
- Logographic
  - Each symbol represents a **word** or **morpheme**, not sounds directly.
  - Chinese characters, Dongba script.
- Morphology: analytic

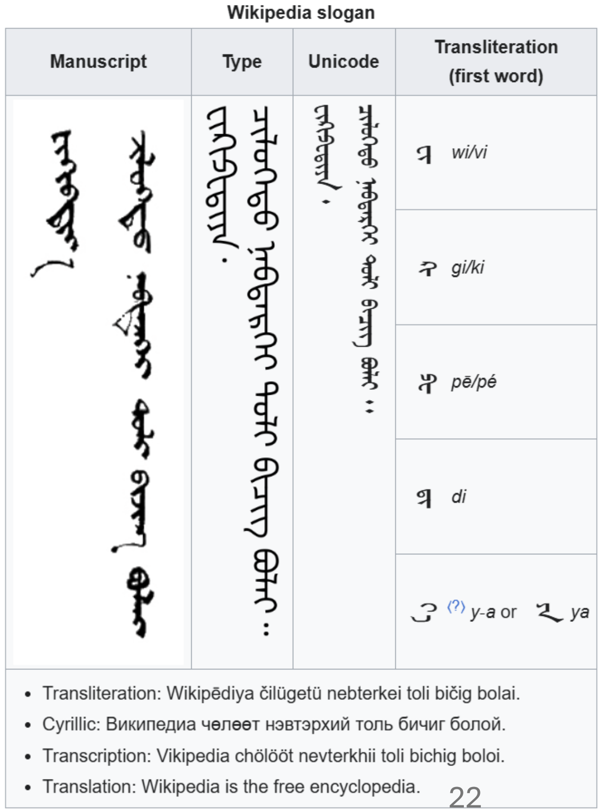
Part of Speech	Chinese Example	English Gloss / Explanation
Verb	**冰**一下可乐。	“Chill the cola for a bit.” → *冰* used as a **verb**, meaning *to make something cold or iced*.
Noun	我爱吃**冰**。	“I love eating ice.” → *冰* as a **noun**, meaning *ice*.
Adjective	可乐真**冰**。	“The cola is really icy.” → *冰* used **adjectivally**, describing temperature or feel (*icy, chilled*).

Major Writing Systems

- Alphabet:
  - A system where each symbol represents a **phoneme**, both consonant and vowel.
  - Latin alphabet, Armenian, Cyrillic, Greek, Georgian, Mongolian, Hangul (Korean).




Աա Բբ Գգ Դդ Եե Զզ Էէ Ըը Թթ Ժժ
a b g d je z e ë t' ž
Իի Լլ Խխ Ծծ Կկ Հհ Ձձ Ղղ Ճճ Մմ
i l x ts k h dz ṙ tš m
Յյ Նն Շշ Ոո Չչ Պպ Ջջ Ռռ Սս Վվ
j n š vo tš' p dž r s v
Տտ Րր Ցց Ււ Փփ Քք Օo Ֆֆ
t r' ts' w p' k' o f



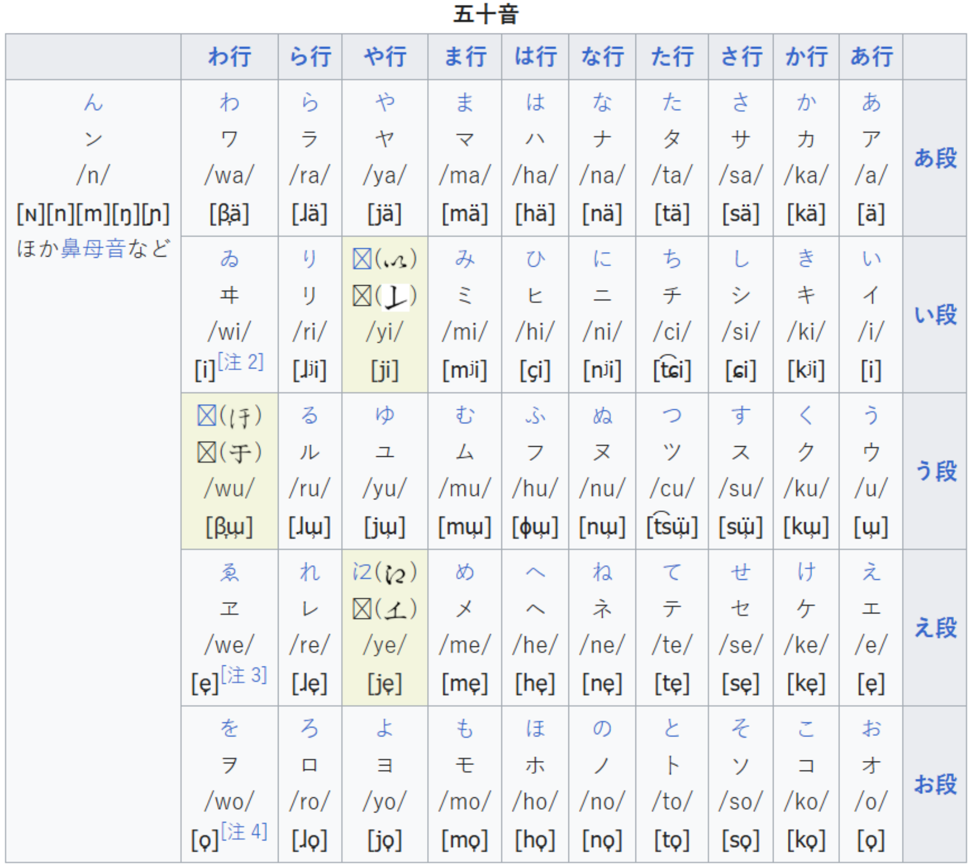
Wikipedia slogan

Manuscript	Type	Unicode	Transliteration (first word)
			wi/vi
			gi/ki
			pē/pé
			di
			(?) y-a or ya

- Transliteration: Wikipēdiya čilügetü nebterkei toli bičig bolai.
- Cyrillic: Википедиа чөлөөт нэвтэрхий толь бичиг болой.
- Transcription: Vikipedia chölööt nevterkhii toli bichig boloi.
- Translation: Wikipedia is the free encyclopedia.

Major Writing Systems

- Syllabary:
  - Each symbol represents a **syllable** (usually CV, CVC, etc.), with no breakdown into phonemes.
  - Cherokee syllabary,
  - Katakana and Hiragana (Japanese)

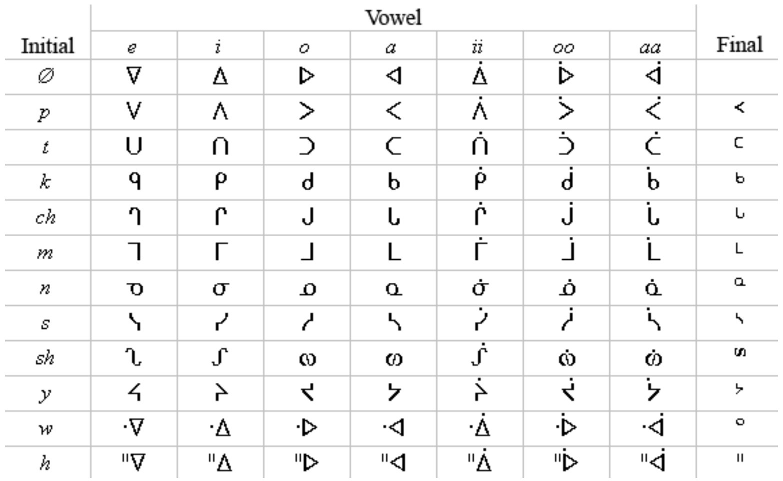


五十音

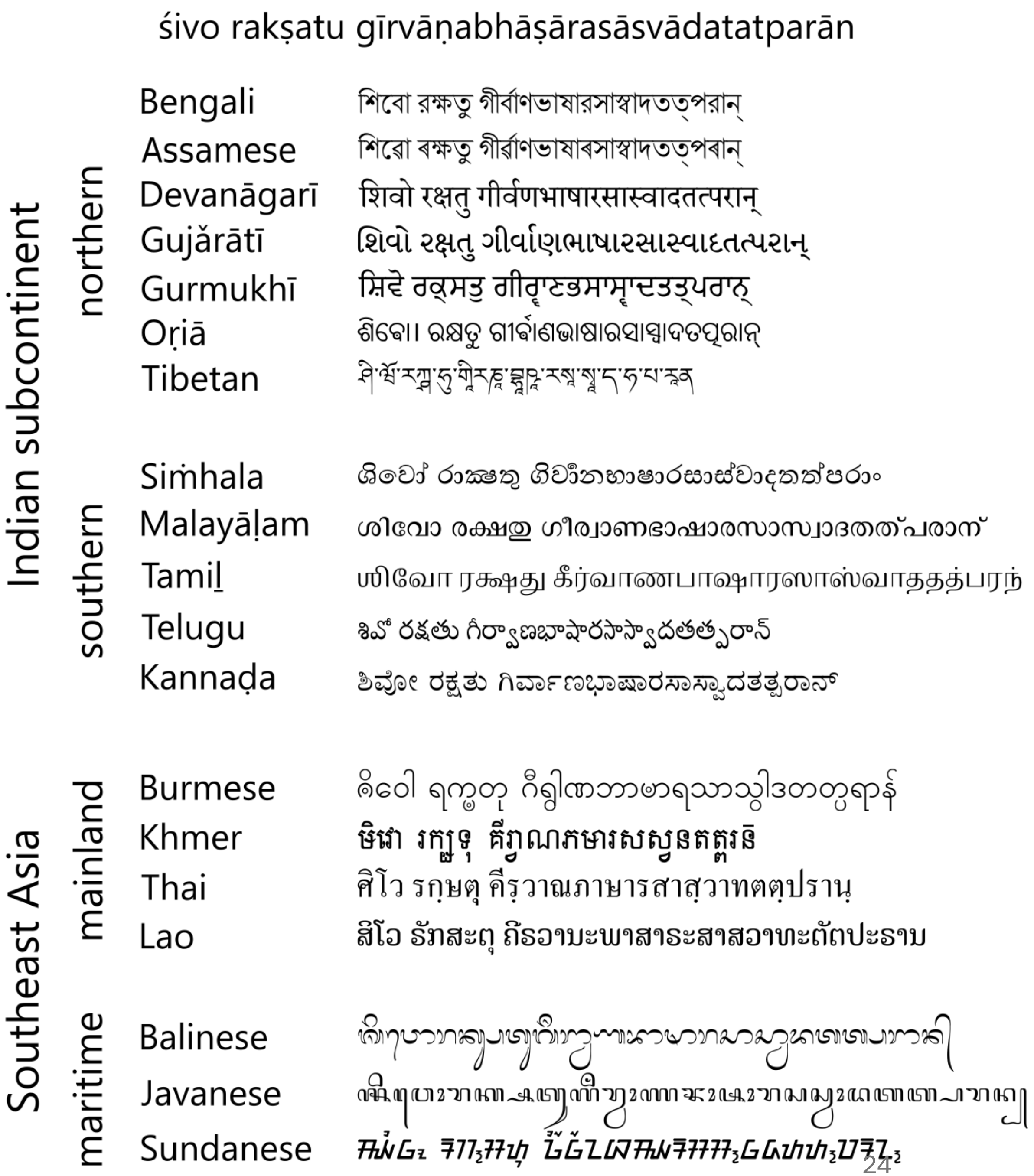
	わ行	ら行	や行	ま行	は行	な行	た行	さ行	か行	あ行	
ん ン /n/ [ɴ][n][m][ŋ][ɲ] ほか鼻母音など	わ ワ /wa/ [β̞ä]	ら ラ /ra/ [ɺä]	や ヤ /ya/ [jä]	ま マ /ma/ [mä]	は ハ /ha/ [hä]	な ナ /na/ [nä]	た タ /ta/ [tä]	さ サ /sa/ [sä]	か カ /ka/ [kä]	あ ア /a/ [ä]	あ段
	ゐ ヰ /wi/ [i][注 2]	り リ /ri/ [ɺʲi]	☒(𛀆) ☒(𛄠) /yi/ [ji]	み ミ /mi/ [mʲi]	ひ ヒ /hi/ [çi]	に ニ /ni/ [nʲi]	ち チ /ci/ [t͡ɕi]	し シ /si/ [ɕi]	き キ /ki/ [kʲi]	い イ /i/ [i]	い段
	☒(𛄟) ☒(𛄢) /wu/ [β̞ɯ̹]	る ル /ru/ [ɺɯ̹]	ゆ ユ /yu/ [jɯ̹]	む ム /mu/ [mɯ̹]	ふ フ /hu/ [ɸɯ̹]	ぬ ヌ /nu/ [nɯ̹]	つ ツ /cu/ [t͡sɯ̹̈]	す ス /su/ [sɯ̹̈]	く ク /ku/ [kɯ̹]	う ウ /u/ [ɯ̹]	う段
	ゑ ヱ /we/ [e̞][注 3]	れ レ /re/ [ɺe̞]	𛀁(𛀁) ☒(𛄡) /ye/ [je̞]	め メ /me/ [me̞]	へ ヘ /he/ [he̞]	ね ネ /ne/ [ne̞]	て テ /te/ [te̞]	せ セ /se/ [se̞]	け ケ /ke/ [ke̞]	え エ /e/ [e̞]	え段
	を ヲ /wo/ [o̞][注 4]	ろ ロ /ro/ [ɺo̞]	よ ヨ /yo/ [jo̞]	も モ /mo/ [mo̞]	ほ ホ /ho/ [ho̞]	の ノ /no/ [no̞]	と ト /to/ [to̞]	そ ソ /so/ [so̞]	こ コ /ko/ [ko̞]	お オ /o/ [o̞]	お段

Major Writing Systems

- Abugida:
  - A system where each symbol represents a **consonant–vowel syllable**, built from a base consonant symbol that can be modified to indicate the vowel.
  - Semitic Ethiopic scripts, Brahmic family of scripts, Canadian Aboriginal syllabics.



Initial	Vowel							Final
	e	*i*	*o*	*a*	*ii*	*oo*	*aa*	
∅	ᐁ	ᐃ	ᐅ	ᐊ	ᐄ	ᐆ	ᐋ	
p	ᐯ	ᐱ	ᐳ	ᐸ	ᐲ	ᐴ	ᐹ	ᑊ
t	ᑌ	ᑎ	ᑐ	ᑕ	ᑏ	ᑑ	ᑖ	ᐟ
k	ᑫ	ᑭ	ᑯ	ᑲ	ᑮ	ᑰ	ᑳ	ᐠ
ch	ᒉ	ᒋ	ᒍ	ᒐ	ᒌ	ᒎ	ᒑ	ᐨ
m	ᒣ	ᒥ	ᒧ	ᒪ	ᒦ	ᒨ	ᒫ	ᒼ
n	ᓀ	ᓂ	ᓄ	ᓇ	ᓃ	ᓅ	ᓈ	ᐣ
s	ᓭ	ᓯ	ᓱ	ᓴ	ᓰ	ᓲ	ᓵ	ᐢ
sh	ᔐ	ᔑ	ᔓ	ᔕ	ᔒ	ᔔ	ᔖ	ᐡ
y	ᔦ	ᔨ	ᔪ	ᔭ	ᔩ	ᔫ	ᔮ	ᔾ
w	ᐌ	ᐎ	ᐒ	ᐗ	ᐐ	ᐔ	ᐙ	ᐤ
h	ᐦᐁ	ᐦᐃ	ᐦᐅ	ᐦᐊ	ᐦᐄ	ᐦᐆ	ᐦᐋ	ᐦ



śivo rakṣatu gīrvāṇabhāṣārasāsvādatatparān

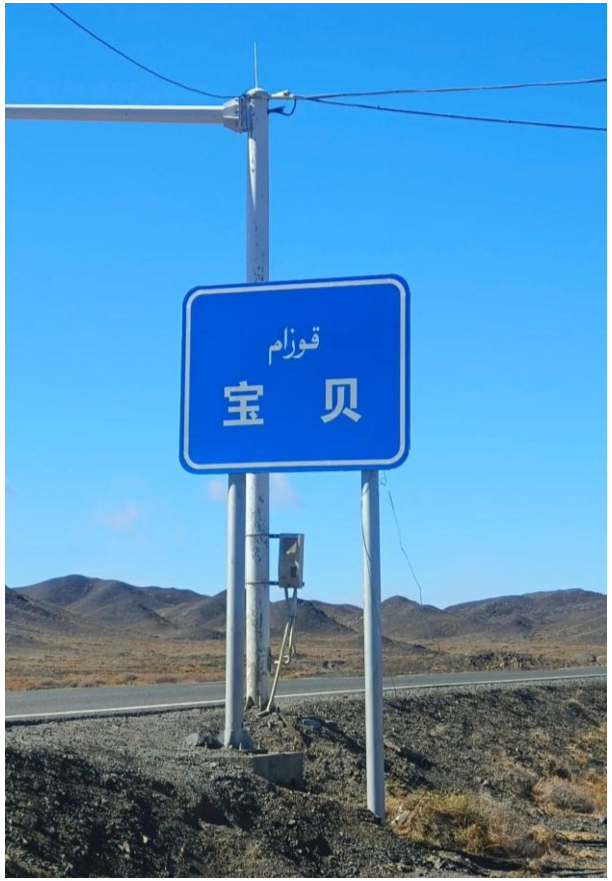
Indian subcontinent	northern	Bengali	শিবো রক্ষতু গীর্বাণভাষারসাস্বাদতত্পরান্
		Assamese	শিৱো ৰক্ষতু গীৰ্ৱাণভাষাৰসাস্বাদতত্পৰান্
		Devanāgarī	शिवो रक्षतु गीर्वणभाषारसास्वादतत्परान्
		Gujărātī	શિવો રક્ષતુ ગીર્વાણભાષારસાસ્વાદતત્પરાન્
		Gurmukhī	ਸ਼ਿਵੋ ਰਕ੍ਸ਼ਤੁ ਗੀਰ੍ਵਾਣਭਾਸ਼ਾਰਸਾਸ੍ਵਾਦਤਤ੍ਪਰਾਨ੍
		Oṛiā	ଶିବୋ। ରକ୍ଷତୁ ଗୀର୍ବାଣଭାଷାରସାସ୍ୱାଦତତ୍ପରାନ୍
		Tibetan	ཤི་ཝོ་རཀྵ་ཏུ་གཱིརྦཱ་ཎ་བྷཱ་ཥཱ་ར་སཱ་སྭཱ་ད་ཏ་པ་རཱན
	southern	Siṁhala	ශිවෝ රාක්‍ෂතු ගිර්වීනභාෂාරසාස්වාදතත්පරාං
		Malayāḷam	ശിവോ രക്ഷതു ഗീര്വാണഭാഷാരസാസ്വാദതത്പരാന്
		Tamil̠	ஶிவோ ரக்ஷது கீர்வாணபாஷாரஸாஸ்வாததத்பரந்
		Telugu	శివో రక్షతు గీర్వాణభాషారసాస్వాదతత్పరాన్
		Kannaḍa	ಶಿವೋ ರಕ್ಷತು ಗಿರ್ವಾಣಭಾಷಾರಸಾಸ್ವಾದತತ್ವರಾನ್
Southeast Asia	mainland	Burmese	ၐိဝေါ ရက္ၑတု ဂီရွါဏဘာၑာရသာသွါဒတတ္ပရာန်
		Khmer	ឝិវោ រក្ឞតុ គីវ៌ាណភាឞារសាស្វាទតត្បរាន៑
		Thai	ศิโว รกฺษตุ คีรฺวาณภาษารสาสฺวาทตตฺปรานฺ
		Lao	ສິໂວ ຣັກສະຕຸ ຄີຣວານະພາສາຣະສາສວາທະຕັຕປະຣານ
	maritime	Balinese	ᬰᬶᬯᭀᬭᬓ᭄ᬱᬢᬸᬕᬷᬃᬯᬵᬡᬪᬵᬱᬭᬲᬵᬲ᭄ᬯᬵᬤᬢᬢ᭄ᬧᬭᬵᬦ᭄
		Javanese	ꦯꦶꦮꦺꦴꦫꦏ꧀ꦰꦠꦸꦒꦷꦂꦮꦴꦟꦨꦴꦰꦫꦱꦴꦱ꧀ꦮꦴꦢꦠꦠ꧀ꦥꦫꦴꦤ꧀
		Sundanese	ᮯᮤᮝᮧ ᮛᮊ᮪ᮯᮒᮥ ᮌᮥᮁᮝᮔᮘᮯᮛᮞᮞ᮪ᮝᮓᮒᮒ᮪ᮕᮛᮔ᮪

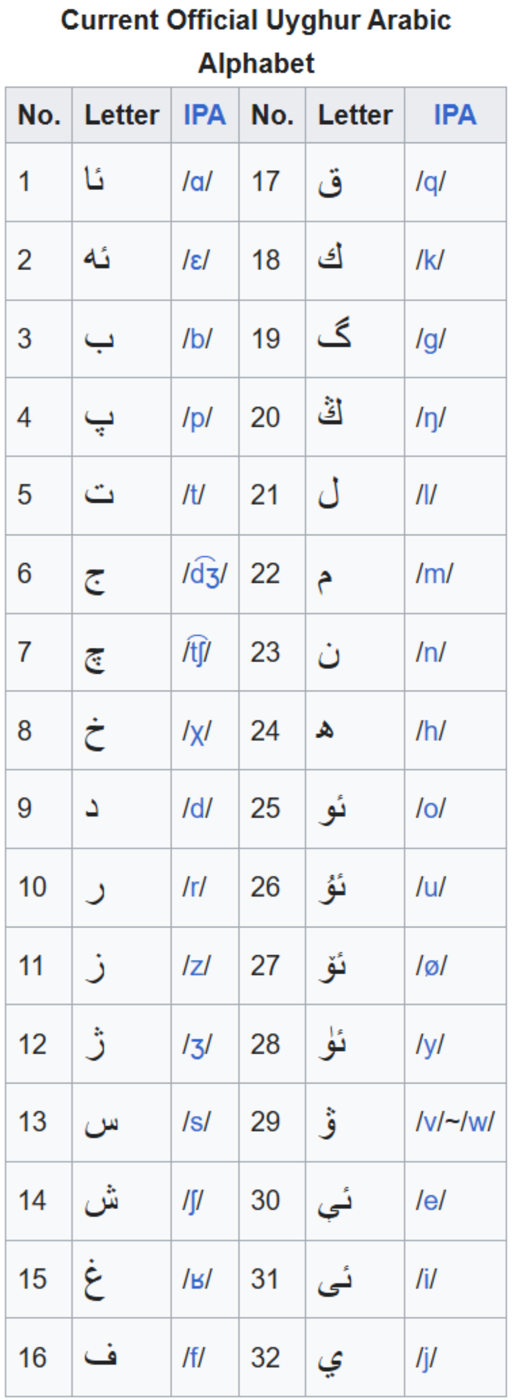
Major Writing Systems

- Abjad:
  - A writing system where each symbol represents a **consonant**, and vowels are usually **omitted** or optional.
  - Arabic and Hebrew.
  - Morphology of semitic languages: triliteral roots
    - Three consonants: root.
    - Vowels: inflectional (and some derivational) changes.
    - K – T – B: scribe, write
    - **k**a**t**a**b**a: wrote (masculine) كَتَبَ كتب
    - **k**u**t**u**b:** books (plural) كُتُب كتب
    - D – R – S: study
    - **d**a**r**a**s**a: studied (masculine)
    - **d**u**r**ū**s:** lessons (plural)

Writing Systems

- Writing systems typically reflect characteristics of the language.
- Categorization of scripts is based on who they are used!
- E.g., modern Uyghur **alphabets**.
- Different tokenization strategy for different writing systems.



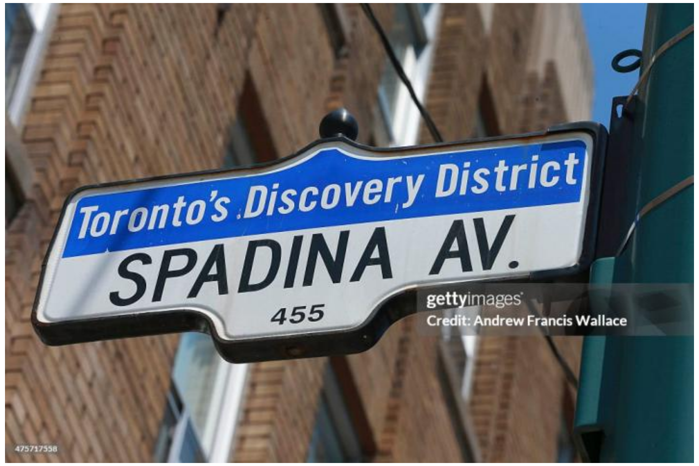


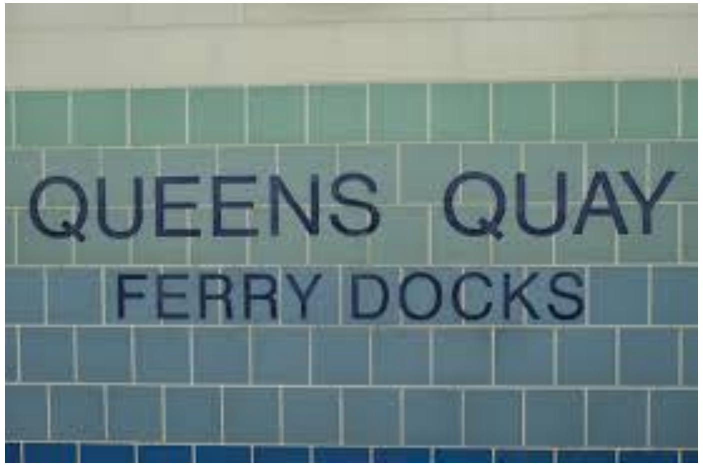
Current Official Uyghur Arabic Alphabet

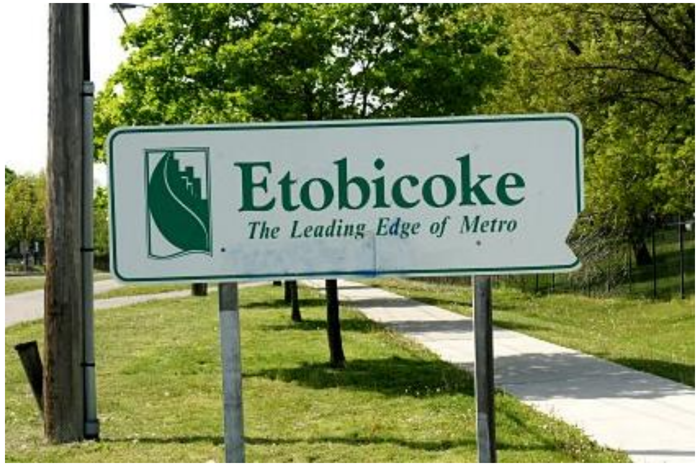
No.	Letter	IPA	No.	Letter	IPA
1	ئا	/ɑ/	17	ق	/q/
2	ئە	/ɛ/	18	ك	/k/
3	ب	/b/	19	گ	/g/
4	پ	/p/	20	ڭ	/ŋ/
5	ت	/t/	21	ل	/l/
6	ج	/d͡ʒ/	22	م	/m/
7	چ	/t͡ʃ/	23	ن	/n/
8	خ	/χ/	24	ھ	/h/
9	د	/d/	25	ئو	/o/
10	ر	/r/	26	ئۇ	/u/
11	ز	/z/	27	ئۆ	/ø/
12	ژ	/ʒ/	28	ئۈ	/y/
13	س	/s/	29	ۋ	/v/~/w/
14	ش	/ʃ/	30	ئې	/e/
15	غ	/ʁ/	31	ئى	/i/
16	ف	/f/	32	ي	/j/

Is English Logographic?

- This is hyperbolic. But spelling does not always accurately indicate pronunciation.
- Receipt, queue, gauge, aisle, debt…
- Toronto place/road names:







Tokenization

- Segmenting an input stream into an ordered sequence of units is called **tokenization**.
- A system which splits texts into tokens is called a **tokenizer.**

A very simple example:

- Input text:
  John likes Mary and Mary likes John.
- Tokens:
  {"John", "likes", "Mary", "and", "Mary", "likes", "John", "."}

Surprisingly hard for English.

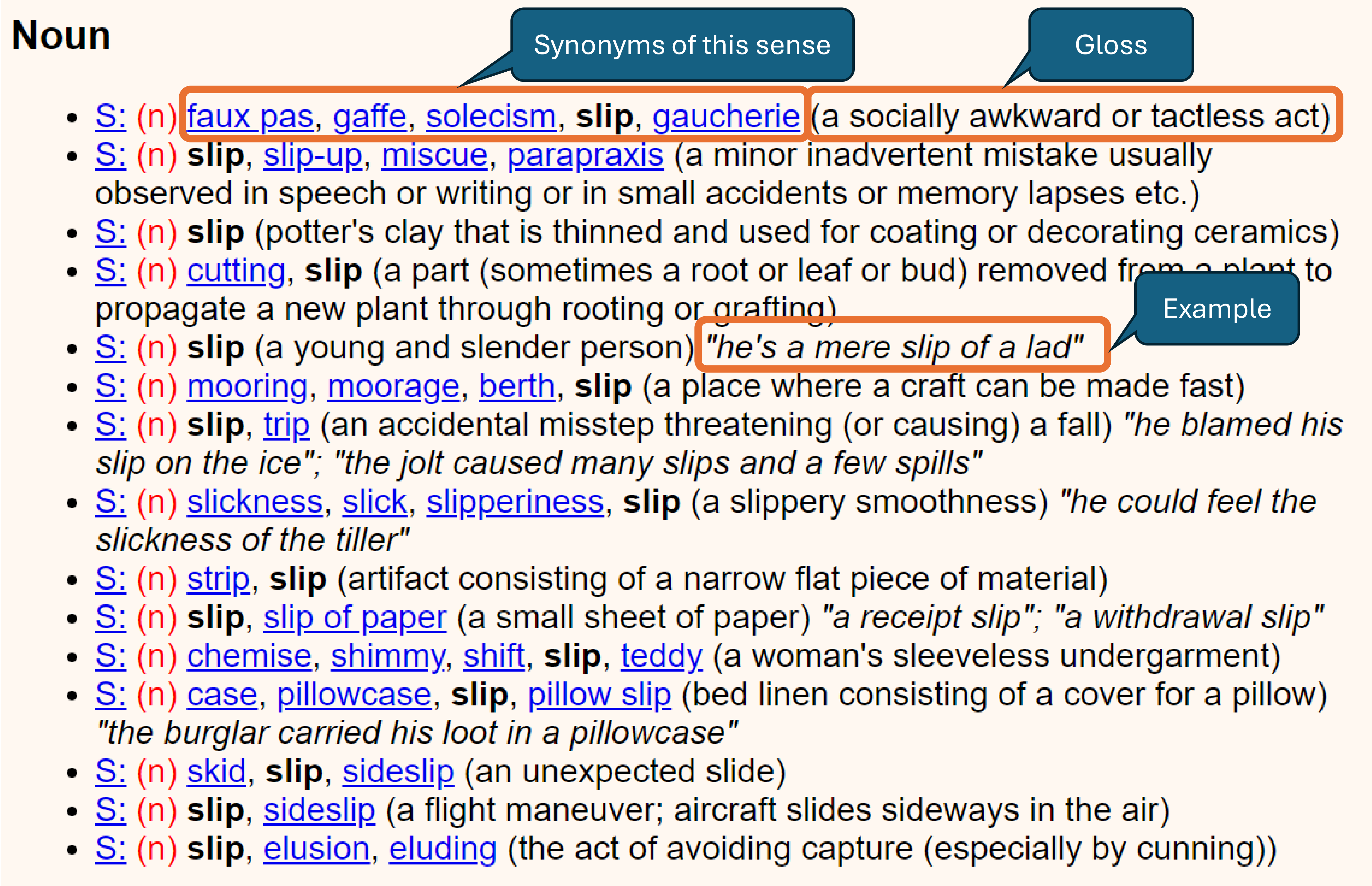
What about semantics?

How to represent “meaning”?

WordNet

- ***WordNet:*** A hierarchical (taxonomic) lexicon and thesaurus of English.
  - Developed by lexicographers at Princeton, 1990s to present.
- Graph structure
  - Nodes are ***synsets*** (“synonym sets”) (≈ word senses).
  - `wn.synsets('dog’)`

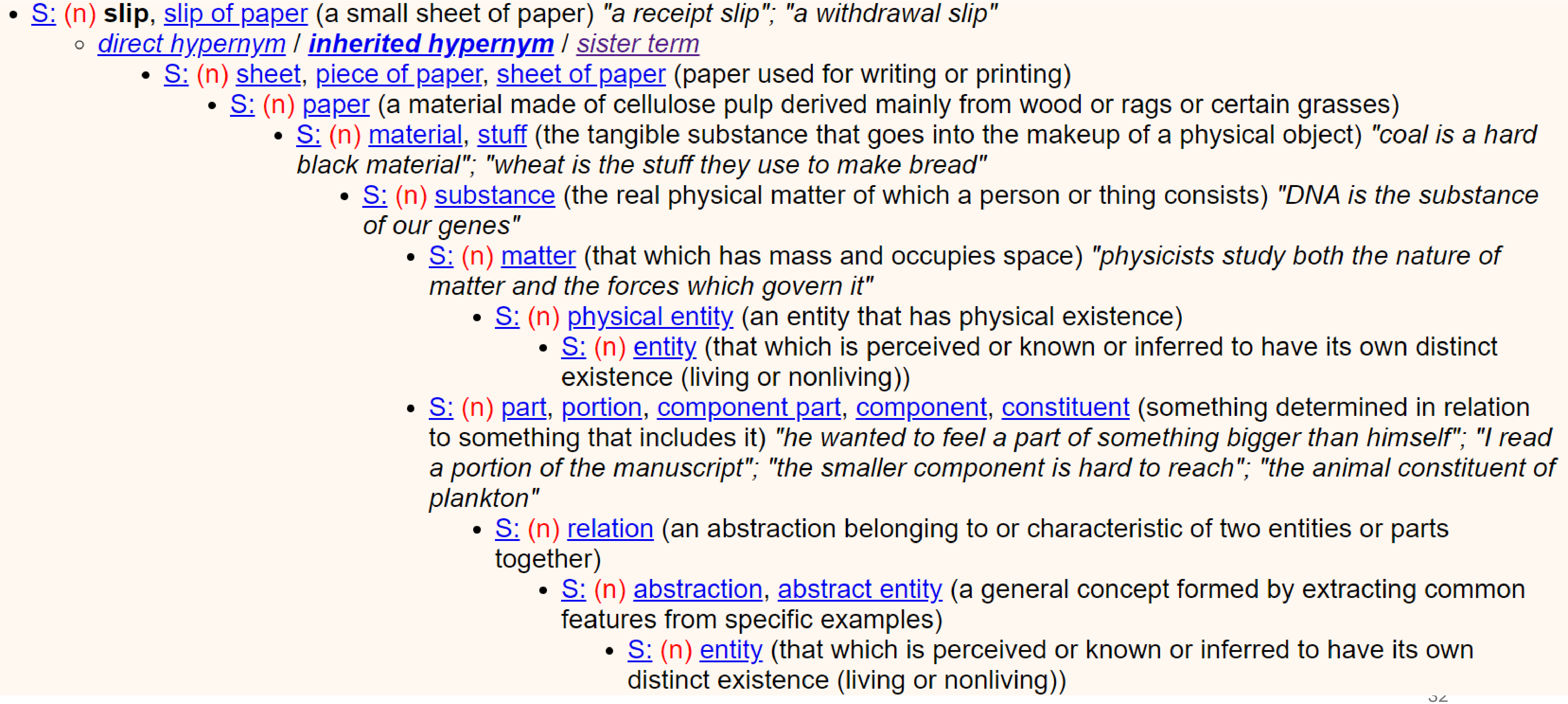
http://wordnetweb.princeton.edu/perl/webwn



Noun

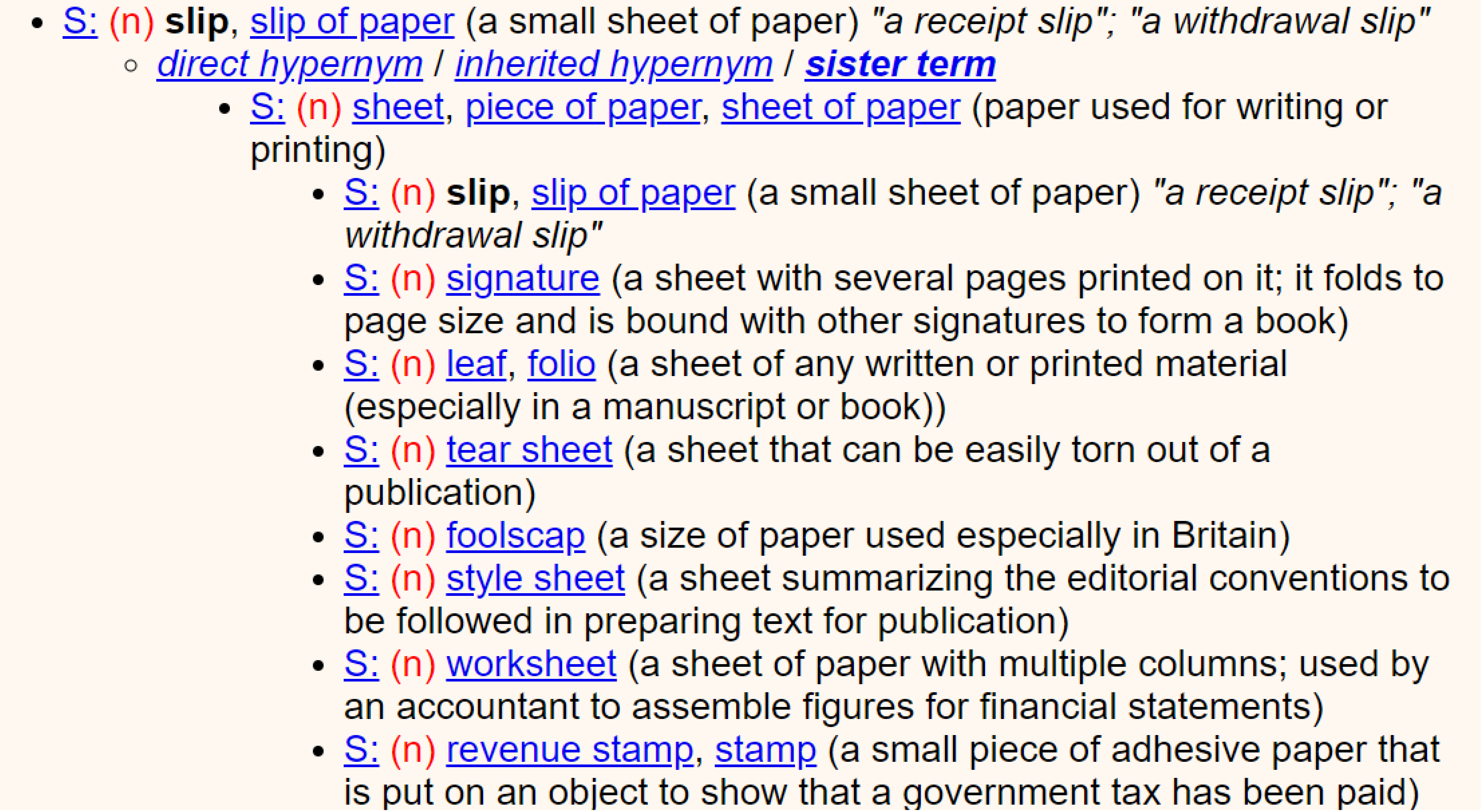
- S: (n) faux pas, gaffe, solecism, **slip**, gaucherie (a socially awkward or tactless act)
- S: (n) **slip**, slip-up, miscue, parapraxis (a minor inadvertent mistake usually observed in speech or writing or in small accidents or memory lapses etc.)
- S: (n) **slip** (potter's clay that is thinned and used for coating or decorating ceramics)
- S: (n) cutting, **slip** (a part (sometimes a root or leaf or bud) removed from a plant to propagate a new plant through rooting or grafting)
- S: (n) **slip** (a young and slender person) *"he's a mere slip of a lad"*
- S: (n) mooring, moorage, berth, **slip** (a place where a craft can be made fast)
- S: (n) **slip**, trip (an accidental misstep threatening (or causing) a fall) *"he blamed his slip on the ice"; "the jolt caused many slips and a few spills"*
- S: (n) slickness, slick, slipperiness, **slip** (a slippery smoothness) *"he could feel the slickness of the tiller"*
- S: (n) strip, **slip** (artifact consisting of a narrow flat piece of material)
- S: (n) **slip**, slip of paper (a small sheet of paper) *"a receipt slip"; "a withdrawal slip"*
- S: (n) chemise, shimmy, shift, **slip**, teddy (a woman's sleeveless undergarment)
- S: (n) case, pillowcase, **slip**, pillow slip (bed linen consisting of a cover for a pillow) *"the burglar carried his loot in a pillowcase"*
- S: (n) skid, **slip**, sideslip (an unexpected slide)
- S: (n) **slip**, sideslip (a flight maneuver; aircraft slides sideways in the air)
- S: (n) **slip**, elusion, eluding (the act of avoiding capture (especially by cunning))

Wordnet hyperonyms

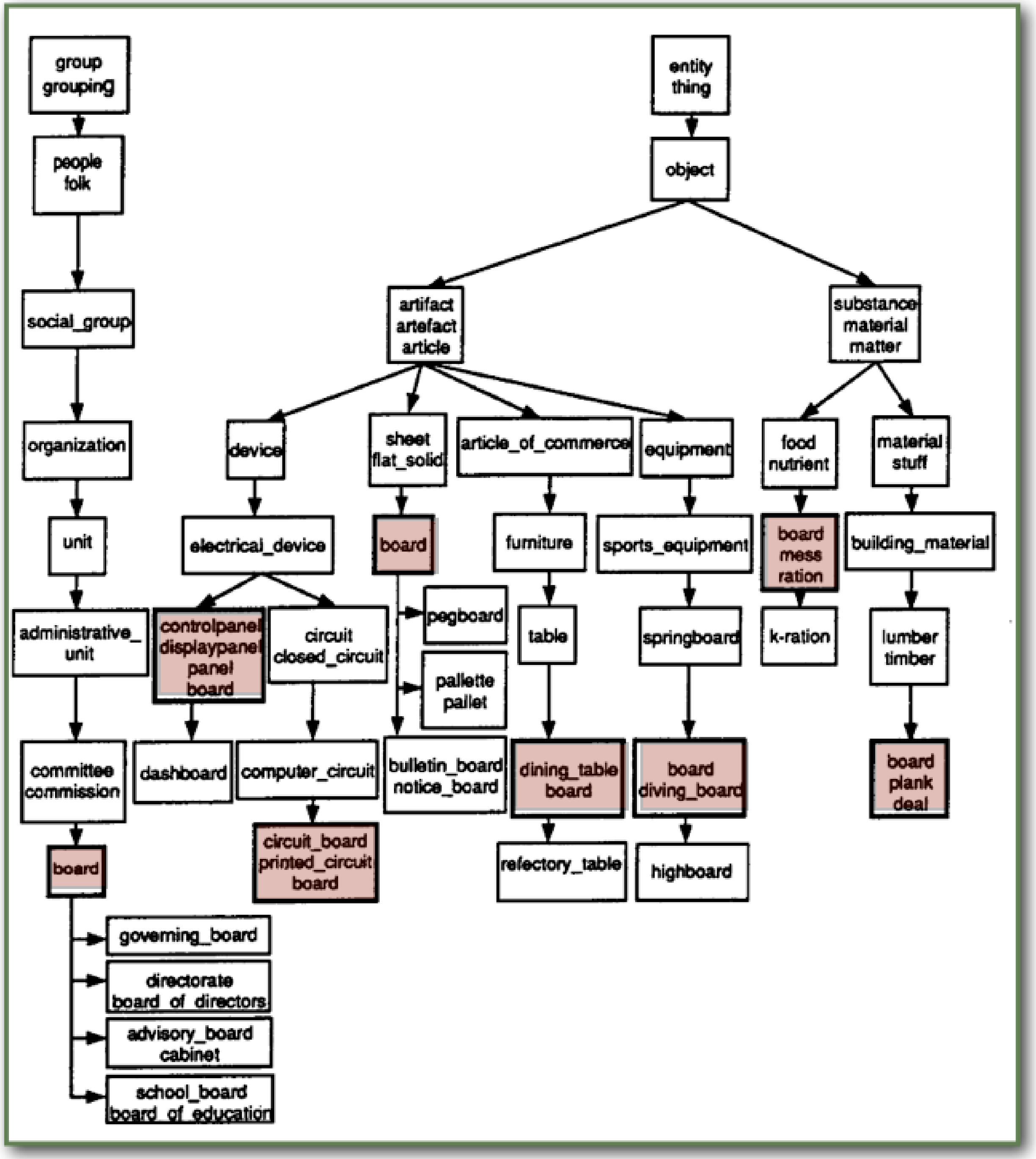


- S: (n) **slip**, slip of paper (a small sheet of paper) *"a receipt slip"; "a withdrawal slip"*
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) sheet, piece of paper, sheet of paper (paper used for writing or printing)
      - S: (n) paper (a material made of cellulose pulp derived mainly from wood or rags or certain grasses)
        - S: (n) material, stuff (the tangible substance that goes into the makeup of a physical object) *"coal is a hard black material"; "wheat is the stuff they use to make bread"*
          - S: (n) substance (the real physical matter of which a person or thing consists) *"DNA is the substance of our genes"*
            - S: (n) matter (that which has mass and occupies space) *"physicists study both the nature of matter and the forces which govern it"*
              - S: (n) physical entity (an entity that has physical existence)
                - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
            - S: (n) part, portion, component part, component, constituent (something determined in relation to something that includes it) *"he wanted to feel a part of something bigger than himself"; "I read a portion of the manuscript"; "the smaller component is hard to reach"; "the animal constituent of plankton"*
              - S: (n) relation (an abstraction belonging to or characteristic of two entities or parts together)
                - S: (n) abstraction, abstract entity (a general concept formed by extracting common features from specific examples)
                  - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Sister terms belong to synsets



- S: (n) **slip**, slip of paper (a small sheet of paper) *"a receipt slip"; "a withdrawal slip"*
    - *direct hypernym* / *inherited hypernym* / ***sister term***
        - S: (n) sheet, piece of paper, sheet of paper (paper used for writing or printing)
            - S: (n) **slip**, slip of paper (a small sheet of paper) *"a receipt slip"; "a withdrawal slip"*
            - S: (n) signature (a sheet with several pages printed on it; it folds to page size and is bound with other signatures to form a book)
            - S: (n) leaf, folio (a sheet of any written or printed material (especially in a manuscript or book))
            - S: (n) tear sheet (a sheet that can be easily torn out of a publication)
            - S: (n) foolscap (a size of paper used especially in Britain)
            - S: (n) style sheet (a sheet summarizing the editorial conventions to be followed in preparing text for publication)
            - S: (n) worksheet (a sheet of paper with multiple columns; used by an accountant to assemble figures for financial statements)
            - S: (n) revenue stamp, stamp (a small piece of adhesive paper that is put on an object to show that a government tax has been paid)



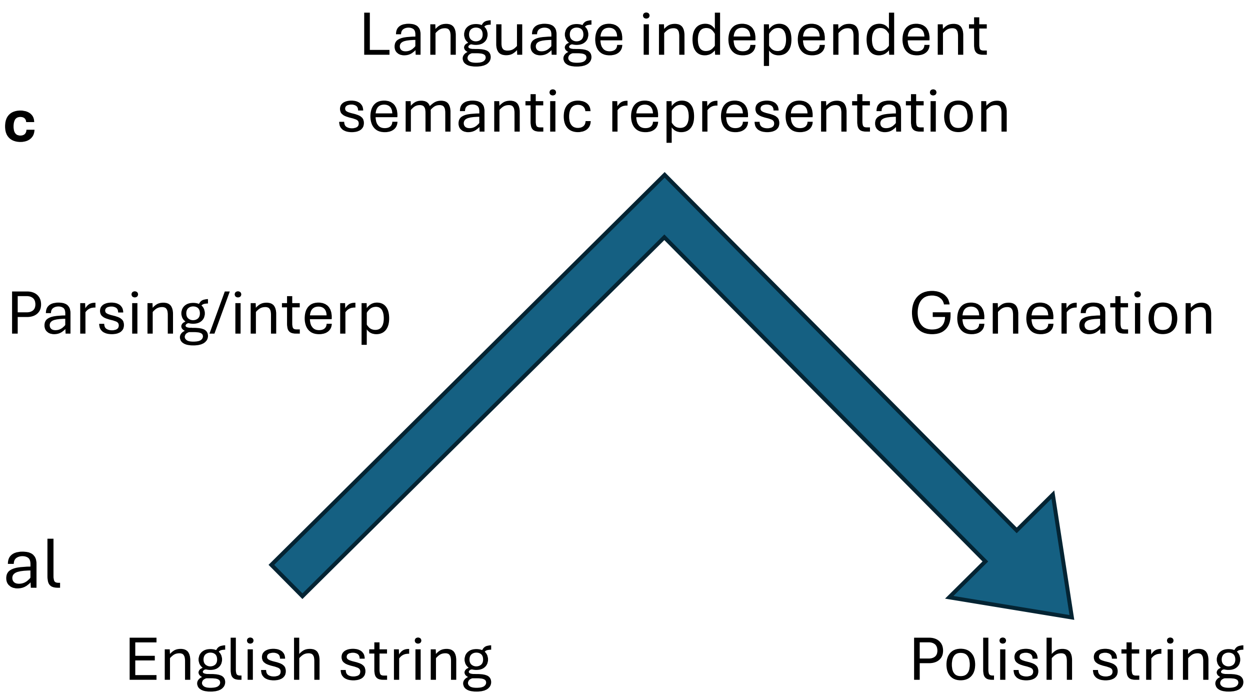
Eight senses of *board* in WordNet, and their hyperonyms and hyponyms.

WordNet

- WordNets now available or under construction for many languages.
  Afrikaans, Albanian, Arabic, Bantu, Basque, Bengali, Bulgarian, Catalan, Chinese, Croatian, Czech, Danish, Dutch, English, Estonian, Farsi (Persian), Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Icelandic, Indonesian, Italian, Irish, Japanese, Kannada, Korean, Latin, Latvian, Macedonian, Maltese, Marathi, Moldavian, Mongolian, Myanmar, Nepali, Norwegian, Oriya, Polish, Portuguese, Romanian, Russian, Sanskrit, Serbian, Slovenian, Spanish, Swedish, Tamil, Thai, Turkish, Vietnamese…
- http://globalwordnet.org/resources/wordnets-in-the-world/

Symbolic NLP

- At the very centre:
  - How do we **represent linguistic information**?
- Earlier attempts:
  - Syntax: parsing
  - Semantics: WordNet
- Next lecture onward: statistical approaches (including neural ones)



Language independent semantic representation

Parsing/interp

Generation

English string

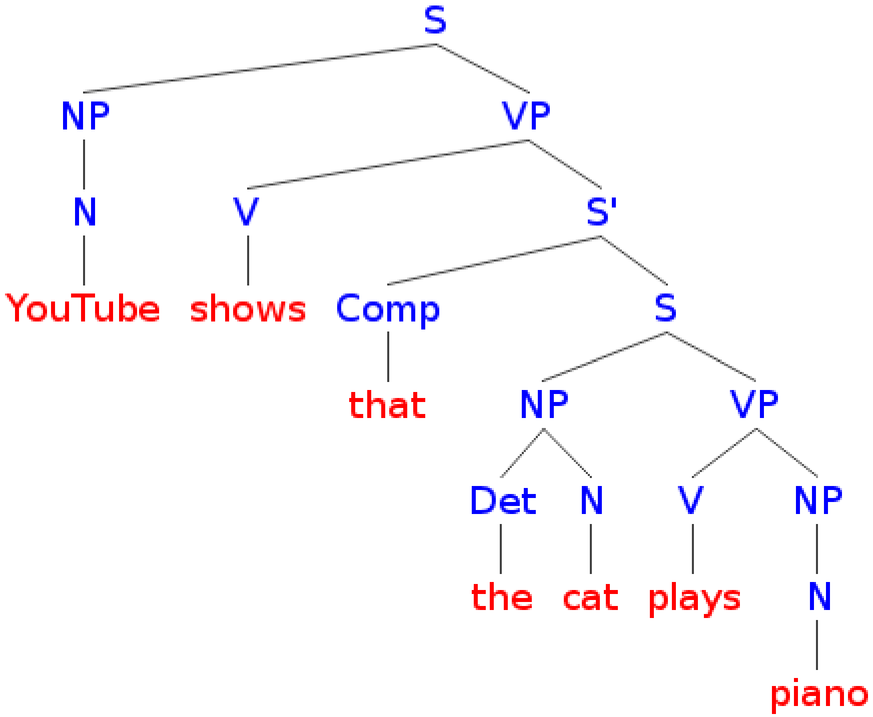
Polish string

Parsing

Elements of Grammar

Primitives

Combinations



Elements of Grammar: *Primitives*

What is *the minimal meaningful units of language*?

- One word or two, or more?
- Ice-cream.
- Donaudampfschifffahrtsgesellschaftskapitän.
- I can read a book.   I read a book yesterday.
- Bass   bass

Fishing for Bass

Elements of Grammar: *Primitives*

- Primitives: lexical categories or parts of speech.
  - Each **word-type** is a member of one or more.
  - Each **word-token** is an instance of exactly one.
    e.g. The cat in the hat sat. *-> 6 tokens, 5 types*.
- Word: *the minimal meaningful units of language*.
- Token: *the minimal units of text*.

Type–token
distinction

Word Categorization

- Categories are **open** or **closed** to new words.
  - *Function word* and *content word*.
- ~~Eight~~ main categories, many subcategories.

~~Nine~~ ~~Seven~~

Twenty-three?

- The categories might possibly be language-specific as well.

Parts of Speech

- **Nouns**: denote an object, a concept, a place, …
    - **Count nouns**: *dog, shoe, Band-Aid, …*
    - **Mass nouns**: *water, wheat, …*
    - **Proper nouns**: *Fred, New York City, …*
- **Pronouns**: *he, she, you, I, they, …*
- **Adjectives**: denote an attribute of the denotation of a noun.
    - Intersective: *pink, furry, …*
    - Measure: *big, small, …*
    - Intensional: *former, alleged, …*

Parts of Speech

- **Verbs**: predicates, denote an action or a state. Numerous distinctions, e.g. transitivity:
  - Intransitive: *sleep, die, …*
  - Transitive: *eat, kiss, …*
  - Ditransitive: *give, sell, …*
  - Copula: *be, feel, become, …*
- **Determiners**, **articles**: specify certain attributes of the denotation of a noun that are grammatically relevant
  - *the, a, some, …*

Parts of Speech

- **Adverbs**: denote an attribute of the denotation of a predicate.
  - Time and place: *today, there, now, …*
  - Manner: *happily, furtively, …*
  - Degree*: *much, very, …*
- **Prepositions**: relate two phrases with a location, direction, manner, etc.
  - *up, at, with, in front of, before, …*

Parts of Speech

- **Conjunctions**: combine two clauses or phrases:
  - Coordinating conjunctions: *and, or, but,* …
  - Subordinating conjunctions: *because, while,* …
- **Interjections**: stand-alone emotive expressions:
  - *um, wow, oh dear,* …

Elements of Grammar: *Combinations*

- ***Combinations***:
  - **Phrase**: a hierarchical grouping of words and/or phrases.
  - **Clause**: a phrase consisting of a verb and (almost) all its dependents.
  - **Sentence**: a clause that is syntactically independent of other clauses.
- Can be represented by tree (or a labelled bracketing).
- Terminology: A ***constituent*** is a well-formed phrase with overtones of semantic and/or psychological significance.

Types of Phrase

- Noun phrase (NP):
  - *a mouse*
  - *mice*
  - *Mickey*
  - *the handsome marmot*
  - *the handsome marmot on the roof*
  - *the handsome marmot whom I adore*

- Verb phrase (VP):
  - *laughed loudly*
  - *give the book to Mary*
  - *slept*
  - *quickly gave the book to Mary*

Types of Phrase

- Adjective phrase (AP, AdjP):
  - *green*
  - *proud of Kyle*
  - *very happy that you went*

- Prepositional phrase (PP):
  - *in the sink*
  - *without feathers*
  - *astride the donkey*

Clauses and Sentences

- Clauses:
  - *Ross remarked upon Nadia’s dexterity*
  - *to become a millionaire by the age of 30*
  - *that her mother had lent her for the banquet*

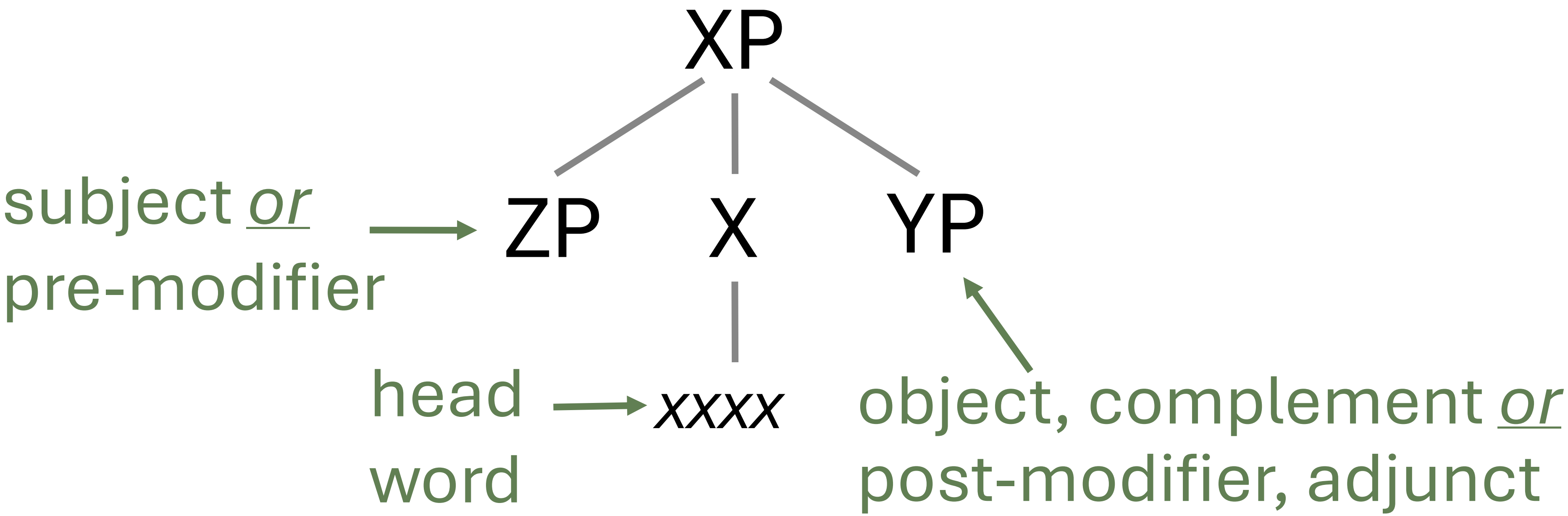
- Sentences:
  - *Ross remarked upon Nadia’s dexterity.*
  - *Nathan wants to become a millionaire by the age of 30.*
  - *Nadia rode the donkey that her mother had lent her for the banquet.*
  - *The handsome marmot on the roof [in dialogue].*

Clauses and Sentences

- Clauses may act as phrases.
- NP:
  - To become a millionaire by the age of 30 is what Ross wants.
  - Nadia riding her donkey is a spectacular sight.
  - Ross discovered that Nadia had been feeding his truffles to the donkey.

The structure of an idealized phrase

XP → ZP X YP



XP

ZP X YP

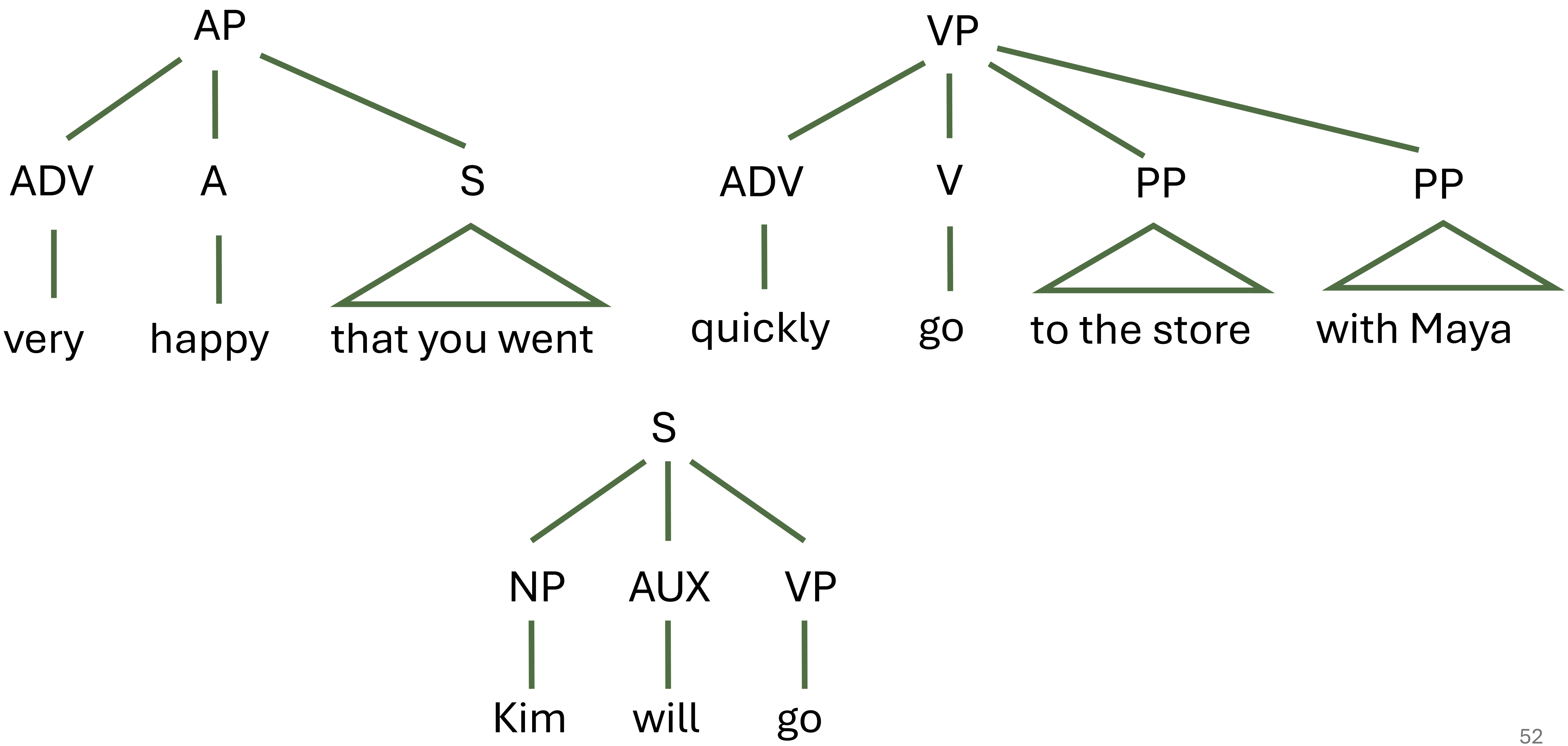
subject *or* pre-modifier

head word

xxxx

object, complement *or* post-modifier, adjunct

Example phrases



S=*S*, *P*={ S → NP VP

NP → Det N

NP → Det Adj N

NP → NP PP

VP → V

VP → V NP

PP → P NP

Det → *the* | *a* | *an*

Adj → *old* | *red* | *happy* | …

N → *dog* | *park* | *ice-cream* | *contumely* | *run* | …

V → *saw* | *ate* | *run* | *disdained* | …

P → *in* | *to* | *on* | *under* | *with* | … }

Lexical categories:
NT's that rewrite
as a single T.

V_t and V_n can be inferred from the production rules.

The lexicon:
In practice, a separate data structure

S=*S*, *P*={ S → NP VP

NP → Det N

NP → Det Adj N

NP → NP PP

VP → V

VP → V NP

PP → P NP

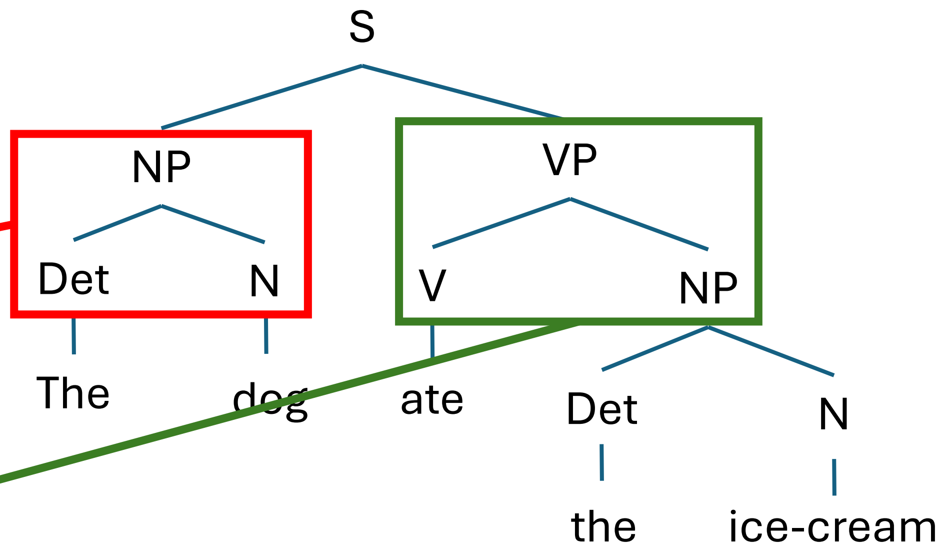
Det → *the* | *a* | *an*

Adj → *old* | *red* | *happy* | …

N → *dog* | *park* | *ice-cream* | *contumely* | *run* | …

V → *saw* | *ate* | *run* | *disdained* | …

P → *in* | *to* | *on* | *under* | *with* | … }



Terminology

- **Non-terminal** (NT):
  A symbol that occurs on the left-hand side (lhs) of some rule.
- **Pre-terminal**:
  a kind of non-terminal located on the LHS of a lexical entry.
- **Terminal** (T):
  A symbol that never occurs on the lhs of a rule.
- **Start symbol**:
  A specially designated NT that must be the root of any tree derived from the grammar.
  In our grammars, it is usually S for sentence.

Parsing

- Parsing: Determining the structure of a sequence of words, given a grammar.
    - Which grammar rules should be used?
    - To which symbols (words / terminals and nodes / non-terminals) should each rule apply?

Parsing

- Input:
  - A context-free grammar.
  - A sequence of words

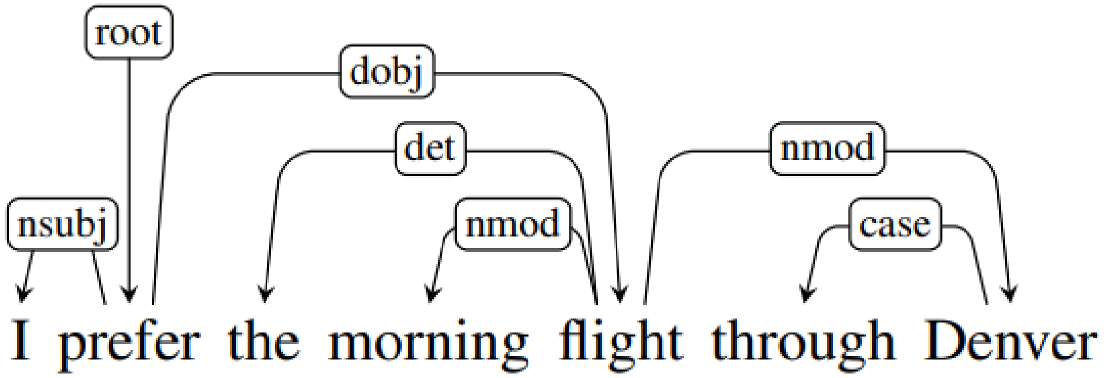
    Time flies like an arrow.

  or, more precisely, of sets of parts of speech.

    {*noun,verb*} {*noun,verb*} {*verb,prep*} {*det*} {*noun*}

- Process:
  - (Working from left to right?,) ***guess*** how each word fits in.

Dependency Grammar



- Arc: from **head** to **dependent**.
- Label on arc: **grammatical function.**
    - Universal Dependencies (UD)
- Dependency vs constituent:
    - More focus on semantics
    - Free word order

Relation	Description	Examples: head -> dep
nsubj	Nominal subject	United canceled the flight.
obj	Direct object	United diverted the flight to Reno.
iobj	Indirect object	We booked her the flight to Miami.
ccomp	Clausal complement	We took the morning flight.
nmod	Nominal modifier	flight to Houston.
amod	Adjectival modifier	Book the cheapest flight
appos	Appositional modifier	United, a unit of UAL, matched the fares.
det	Determiner	The flight was canceled.
…	…	…

Word Dependency Parsing

Raw sentence

He reckons the current account deficit will narrow to only 1.8 billion in September.

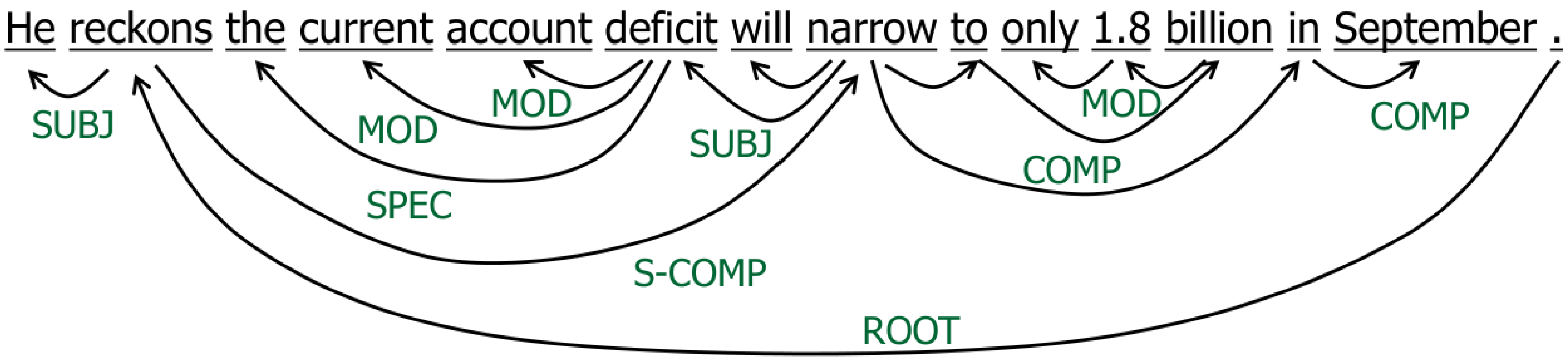
↓ Part-of-speech tagging

POS-tagged sentence

He	reckons	the	current	account	deficit	will	narrow	to	only	1.8	billion	in	September	.
PRP	VBZ	DT	JJ	NN	NN	MD	VB	TO	RB	CD	CD	IN	NNP	.

↓ Word dependency parsing

Word dependency parsed sentence


He reckons the current account deficit will narrow to only 1.8 billion in September .
SUBJ
MOD
MOD
SPEC
SUBJ
S-COMP
MOD
COMP
COMP
ROOT

Dependency Graphs

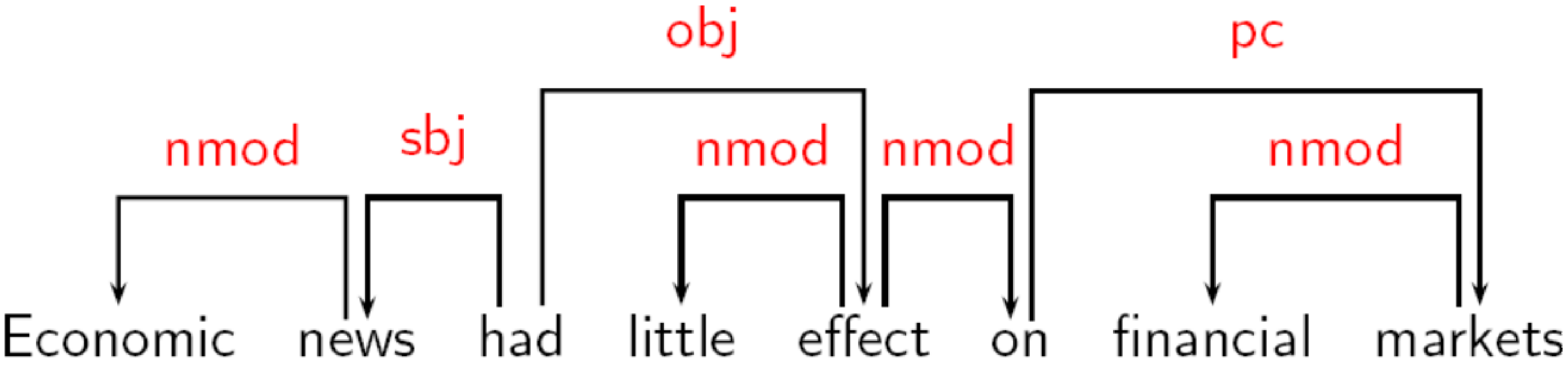
- A dependency structure can be defined as a directed graph G:
  - A set V of nodes,
  - A set E of arcs (edges),
  - A linear precedence order < on V.
- Labelled graphs:
  - Nodes in V are labelled with word forms (and annotation).
  - Arcs in E are labelled with dependency types.
- Notational conventions (i, j ∈ V):
  - $i \rightarrow j \equiv (i, j) \in E$
  - $i \rightarrow^* j \equiv i = j \vee \exists k : i \rightarrow k, k \rightarrow^* j$

Formal Conditions on Dependency Graphs

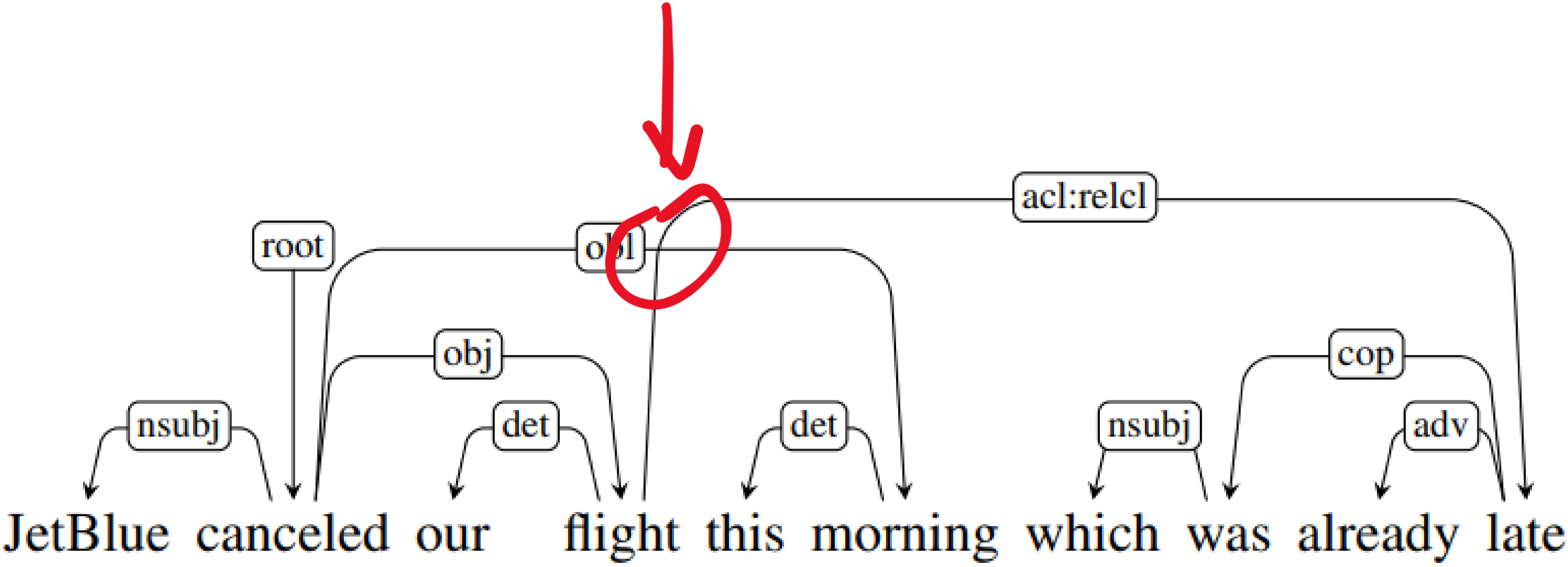
- G is (weakly) connected:
  - For every node i, there is a node j such that $i \to j$ or $j \to i$.
- G is acyclic:
  - If $i \to j$ then not $j \to^* i$.
- G obeys the single-head constraint:
  - If $i \to j$, then not $k \to j$, for any $k \neq i$.
- G is projective:
  - If $i \to j$ then $i \to^* k$, for any k such that $i < k < j$ or $j < k < i$.
  - *No crossing edges*.

Connectedness, Acyclicity and Single-Head

- Intuitions:
  - Syntactic structure is complete. [Connectedness]
  - Syntactic structure is hierarchical. [Acyclicity]
  - Every word has at most one syntactic head. [Single Head]
- Connectedness can be enforced by adding a special ROOT node.



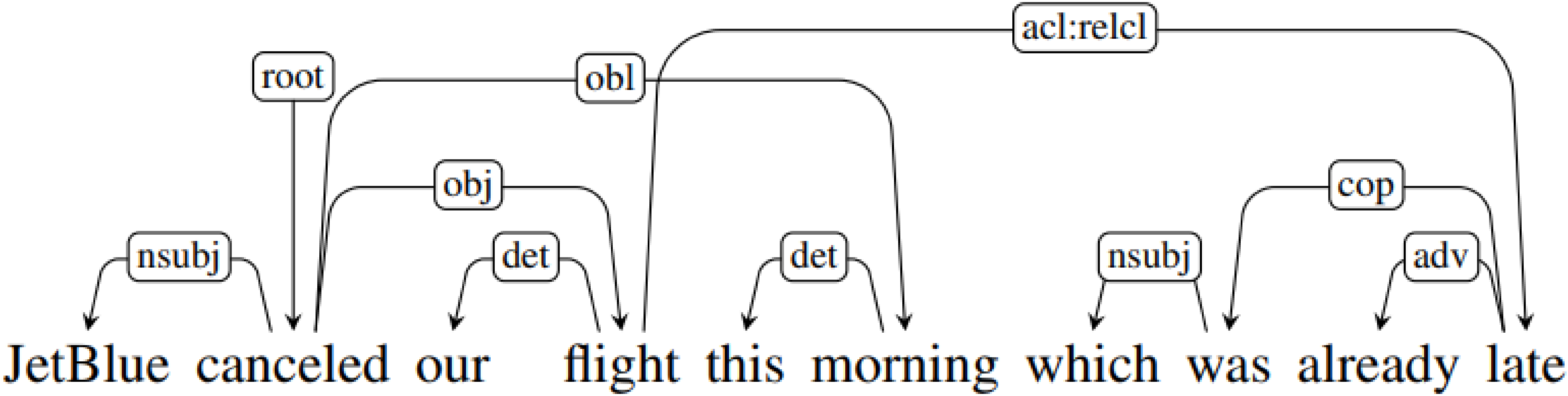
Projectivity



acl:relcl
root
obl
obj
nsubj
det
det
nsubj
cop
adv
JetBlue canceled our flight this morning which was already late

Projectivity

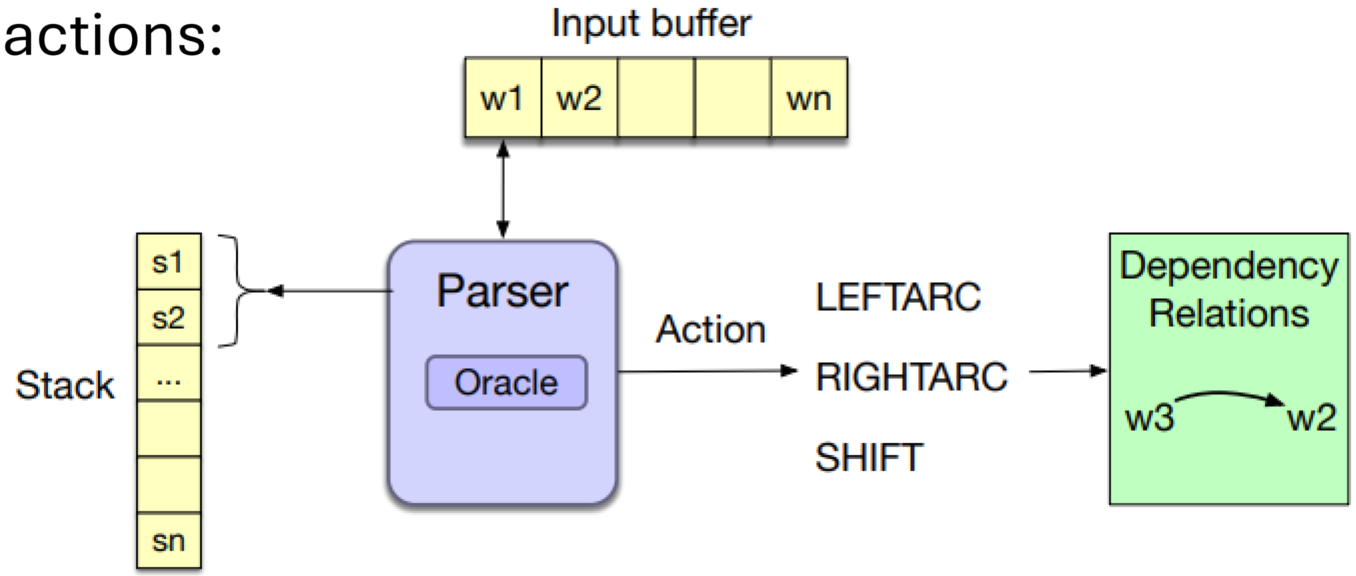
- Most theoretical frameworks do not assume projectivity.
- Non-projective structures are needed to account for:
  - Long-distance dependencies,
  - Free word order.


acl:relcl
root
obl
obj
cop
nsubj
det
det
nsubj
adv
JetBlue canceled our flight this morning which was already late

Transition-Based Dependency Parsing

Shift-Reduce Parsing:

- Data structures:
  - Stack: $[\ldots, w_i]_S$ of partially processed tokens
  - Queue: $[w_j, \ldots]_Q$ of remaining input tokens.
- Parsing actions built from atomic actions:
  - Adding arcs: $(w_i \rightarrow w_j,\ w_i \leftarrow w_j)$.
  - Stack and queue operations.
- Left-to-right parsing in O(n) time.
- Restricted to projective dependency graphs.
  - Non-projective: next week.



Yamada’s Algorithm

- Tree parsing actions:

$$\text{Shift}\qquad \frac{[\ldots]_S \quad [w_i, \ldots]_Q}{[\ldots, w_i]_S \quad [\ldots]_Q}$$

$$\text{Right}\qquad \frac{[\ldots, w_i, w_j]_S \quad [\ldots]_Q}{[\ldots, w_i]_S \quad [\ldots]_Q \quad w_i \longrightarrow w_j}$$

$$\text{Left}\qquad \frac{[\ldots, w_i, w_j]_S \quad [\ldots]_Q}{[\ldots, w_j]_S \quad [\ldots]_Q \quad w_i \longleftarrow w_j}$$

- Algorithm variants:
  - Originally developed for Japanese (strictly head-final) with only the Shift and Left actions [Kudo and Matsumoto 2002].
  - Adapted for English (with mixed headedness) by adding the Left action [Yamada and Matsumoto 2003].

Example

Stack: [root]
Queue:[Book, me, the, morning, flight]

$[\texttt{root}]_S$ $[\texttt{Book me the morning flight}]_Q$

Example

Stack: [root, Book]
Queue:[me, the, morning, flight]

$[\texttt{root Book}]_S$ $[\texttt{me the morning flight}]_Q$

Shift

Example

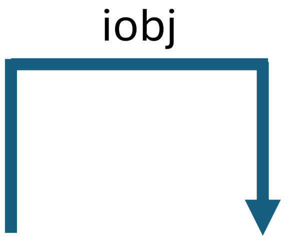
Stack: [root, Book, me]
Queue:[the, morning, flight]

[root Book me]$_S$ [the morning flight]$_Q$

Shift

Example

Stack: [root, Book]
Queue:[the, morning, flight]

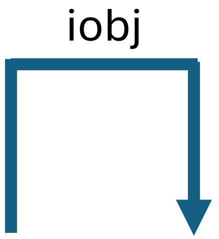


[root Book]$_S$ me [the morning flight]$_Q$

Right

Example

Stack: [root, Book, the]
Queue:[morning, flight]

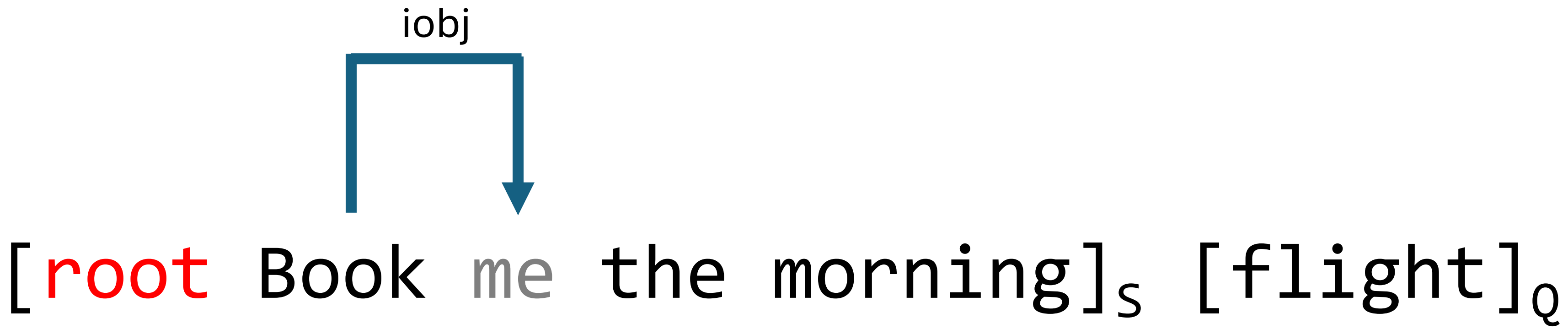


[root Book me the]$_S$ [morning flight]$_Q$

Shift

Example

Stack: [root, Book, the, morning]
Queue:[flight]

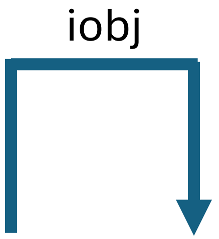


[root Book me the morning]$_S$ [flight]$_Q$

Shift

Example

Stack: [root, Book, the, morning, flight]
Queue:[]

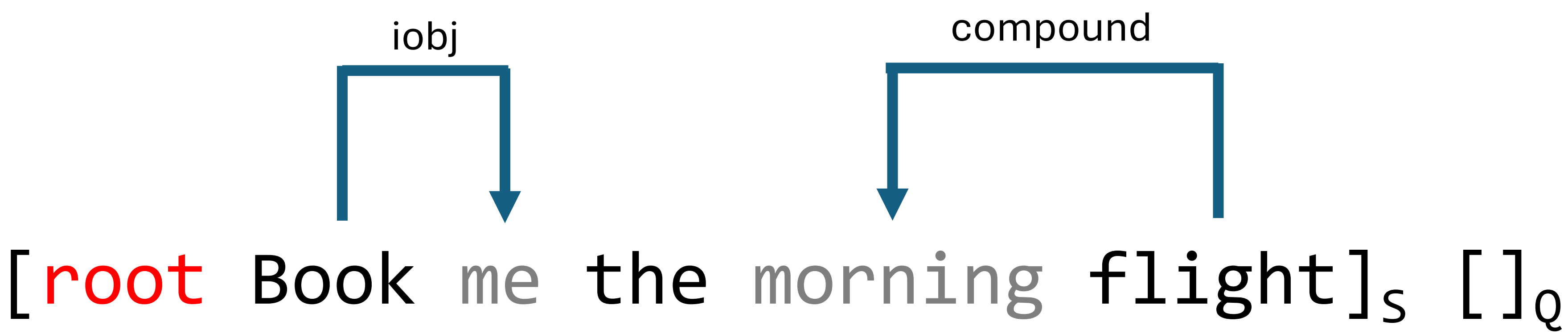


[root Book me the morning flight]$_S$ []$_Q$

Shift

Example

Stack: [root, Book, the, flight]
Queue:[]

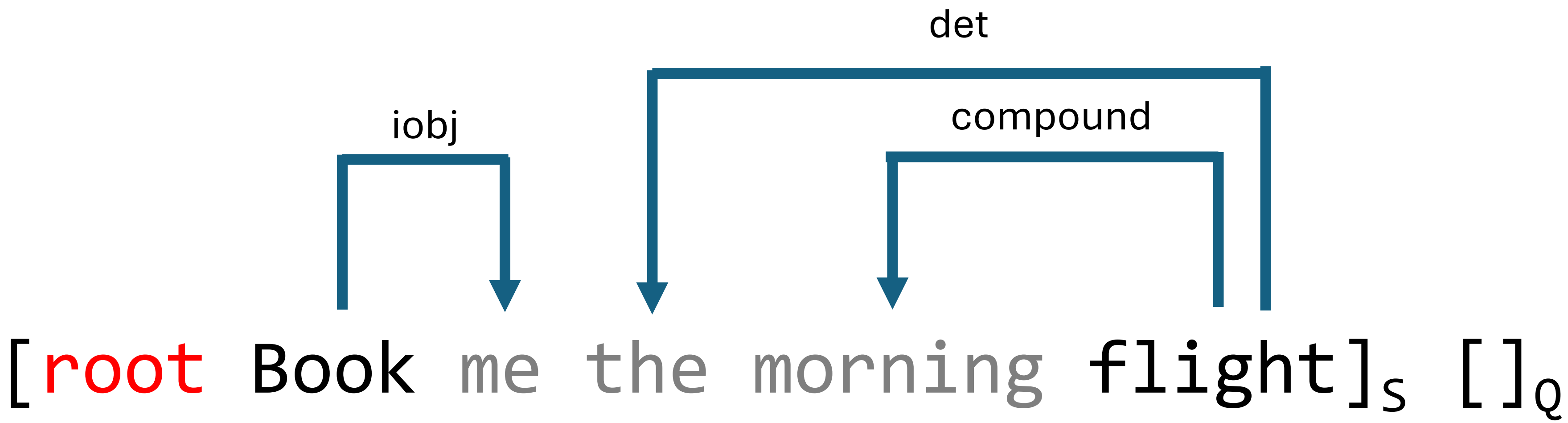


[root Book me the morning flight]$_S$ []$_Q$

Left

Example

Stack: [root, Book, flight]
Queue:[]

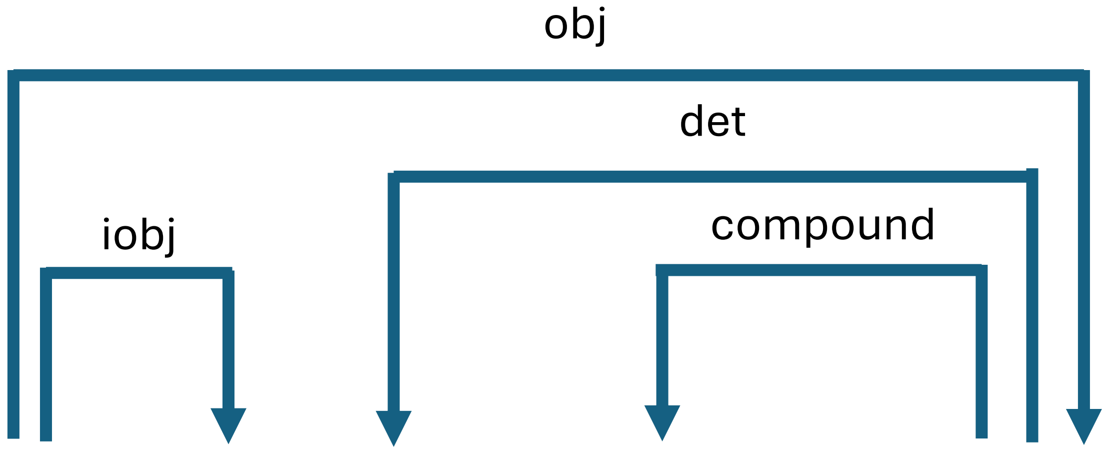


[root Book me the morning flight]$_S$ []$_Q$

Left

Example

Stack: [root, Book]
Queue:[]



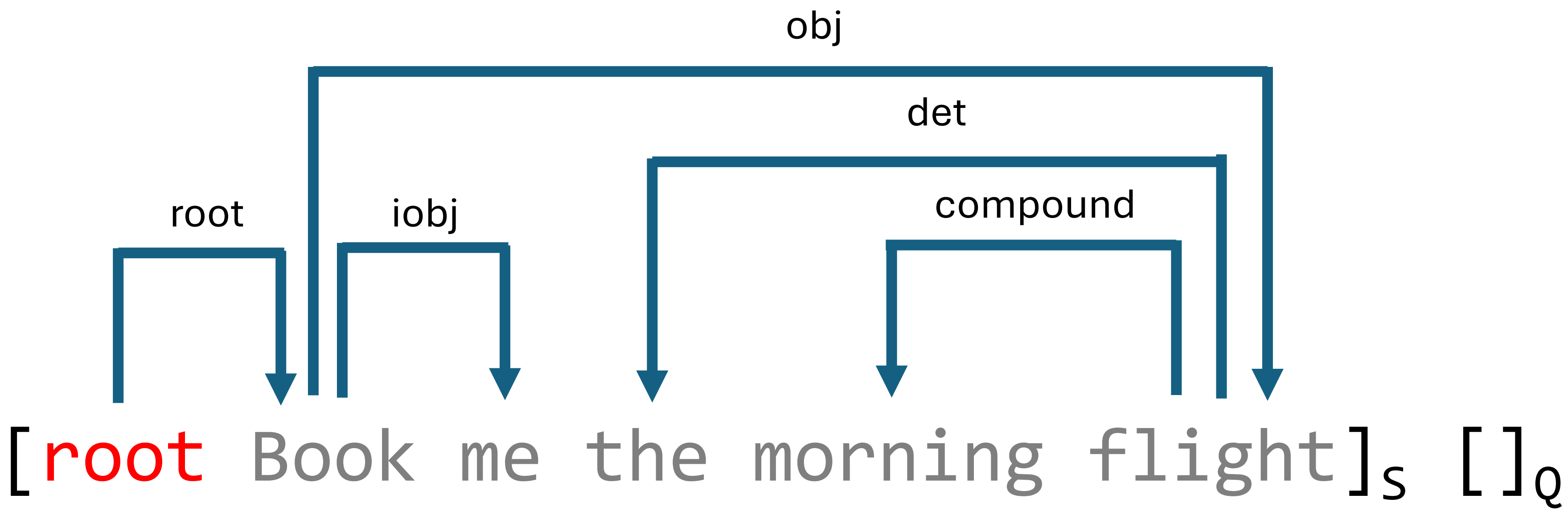
[root Book me the morning flight]$_S$ []$_Q$

Right

Example

Stack: [root]

Queue:[ ]

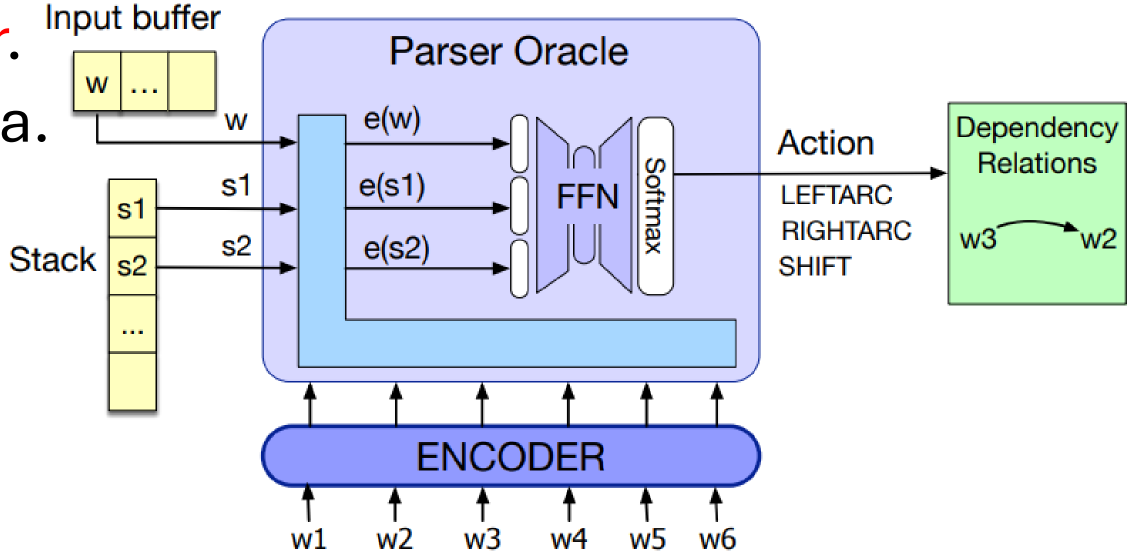


[root Book me the morning flight]$_S$ []$_Q$

Done!

Classifier-Based Parsing

- Data-driven deterministic parsing:
  - Deterministic parsing requires an oracle.
  - An oracle can be approximated by a classifier.
  - A classifier can be trained using treebank data.
- Learning methods:
  - Support vector machines (SVM)
    [Kudo and Matsumoto 2002, Yamada and Matsumoto 2003, Isozaki et al. 2004, Cheng et al. 2004, Nivre et al. 2006]
  - Memory-based learning (MBL)
    [Nivre et al. 2004, Nivre and Scholz 2004]
  - Maximum entropy modelling (MaxEnt)
    [Cheng et al. 2005]
  - Neural networks
    [You! 2024] A1



Ambiguity and Word Sense Disambiguation

Ambiguity!

- Ambiguity at all levels.
    - Lexical
    - Syntactic
    - Semantic
    - Pragmatic

Lexical Ambiguity: Homonymy

- ***Homonymy***: meanings are unrelated.
  [Etymology or history of word is not a deciding factor.]
- Due to same spelling (*homography*):
  - *bank* for money, *bank* of river, *bank* of switches,
    … *bank* → *banque* or *bord* or *rangée* or …?
  - *bass*: “bȧss” fish, “bāss” guitar;
    bow: “bau” to the audience, tie a “bō”.
- Due to same sound (homophony):
  - *wood, would; weather, whether; you, ewe, yew; bough, bow.*

Lexical ambiguity: Polysemy

- ***Polysemy:*** meanings are related.
  - *run*: of humans, rivers, buses, bus routes, …
    line: of people, of type, drawn on paper, transit, route, …

- Often, no clear line between polysemy and homonymy.

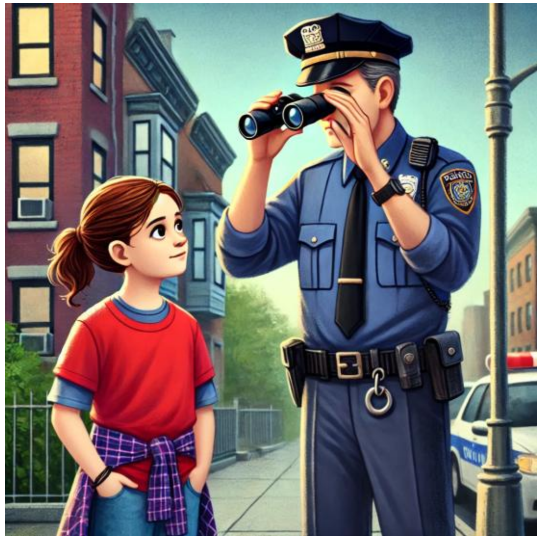
Lexical ambiguity: Polysemy

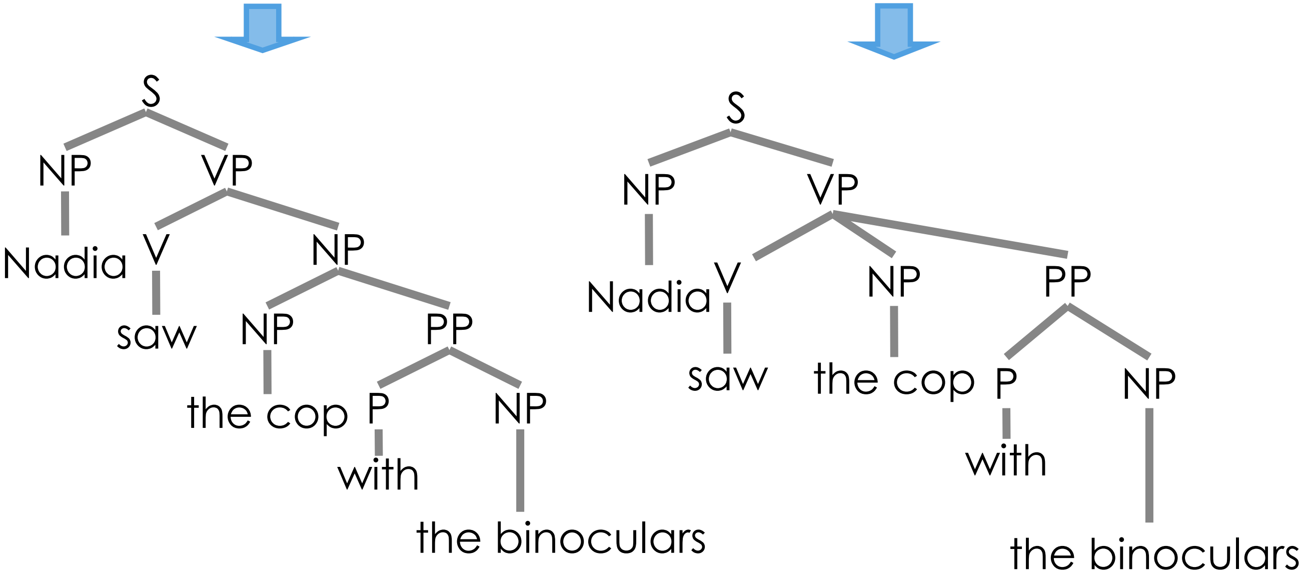
- Sense modulation by context:
  - fast train, fast typist, fast road.
- Systematic polysemy or sense extension:
  - Arrive
    - to come to locations *arrive at the gate*
    - to come to an event *arrive at a concert*
    - to achieve a goal or cognitive state *arrive at a conclusion*
  - Applies to most or all senses of certain semantic classes.

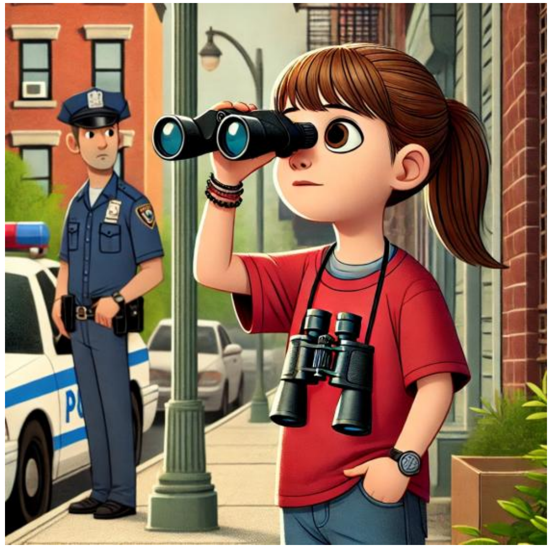
Yu and Xu. Word Sense Extension. ACL 2023.

Syntactic Ambiguity

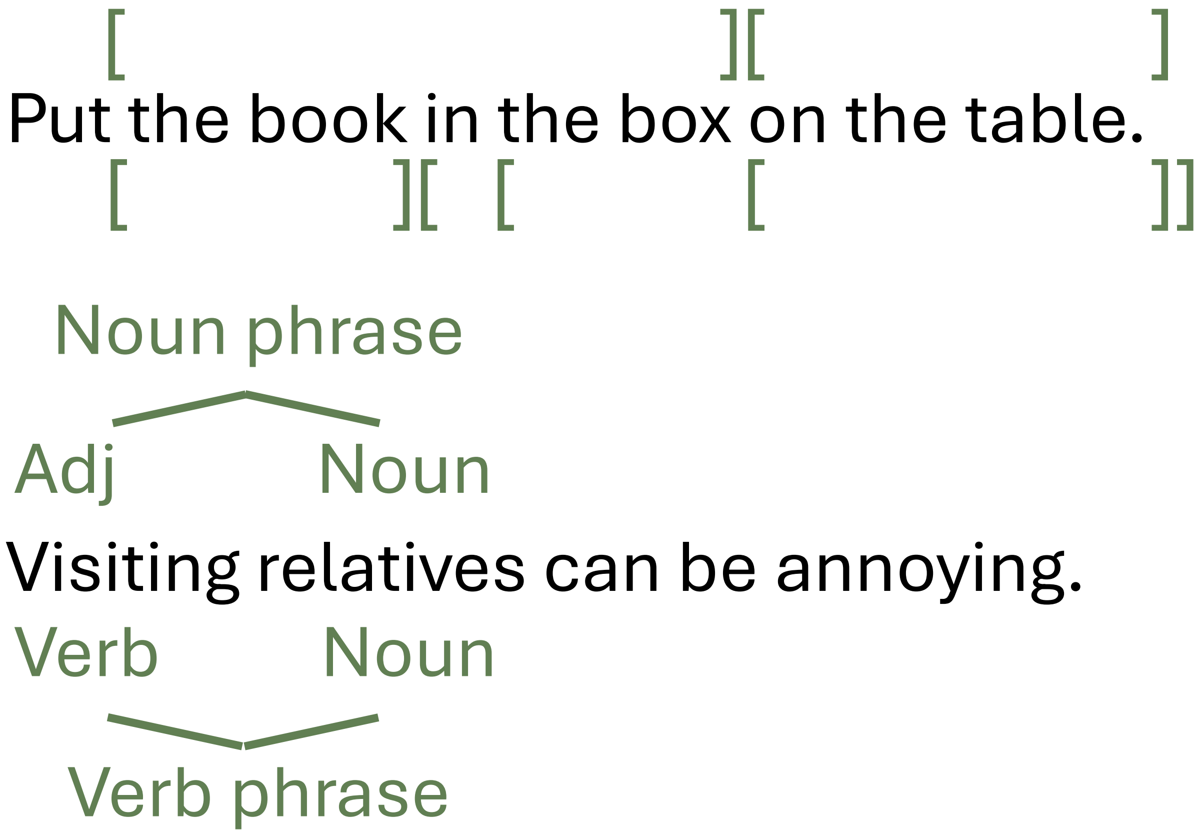
Nadia saw the cop with the binoculars.







Syntactic Ambiguity



Put the book in the box on the table.

Visiting relatives can be annoying.

Syntactic Ambiguity

- These are absolutely everywhere. Some real headlines:
  - *Juvenile Court to Try Shooting Defendant*
  - *Teacher Strikes Idle Kids*
  - *Stolen Painting Found by Tree*
  - *Clinton Wins on Budget, but More Lies Ahead*
  - *Hospitals are Sued by 7 Foot Doctors*
  - *Ban on Nude Dancing on Governor's Desk*
- Usually we don't even notice – we're that good at this kind of resolution.

Syntactic Ambiguity

- Most syntactic ambiguity is **local** — resolved by syntactic or semantic context.
  Visiting relatives is trying.
  Visiting relatives are trying.
  Nadia saw the cop with the gun.
- Sometimes, resolution comes too fast!

  [ ][ ][[????
  The cotton clothing is made from comes from Mississippi.
  [[ ][ ]][[ ]]

  “Garden-path” sentences.

Semantic Ambiguity

- Sentence can have more than one meaning, even when the words and structure are agreed on.
  - *Nadia wants a dog like Ross's*.
  - *Everyone here speaks two languages*.
  - *Iraqi Head Seeks Arms*.
  - *Darmstadt Undergrads Make Nutritious Snacks*.

Pragmatic Ambiguity

- A sample dialogue
  - Nadia: Do you know who’s going to the party?
    Emily: Who?
    Nadia: I don’t know.
    Emily: Oh … I think Carol and Amy will be there.

Quiz

- A lawyer approached the bar.
- The chicken is ready to eat.
- Do you know what time it is?
- Squad helps dog bite victim.
- Miners refuse to work after death.

- Lexical
- Semantic
- Pragmatic
- Syntactic
- Semantic

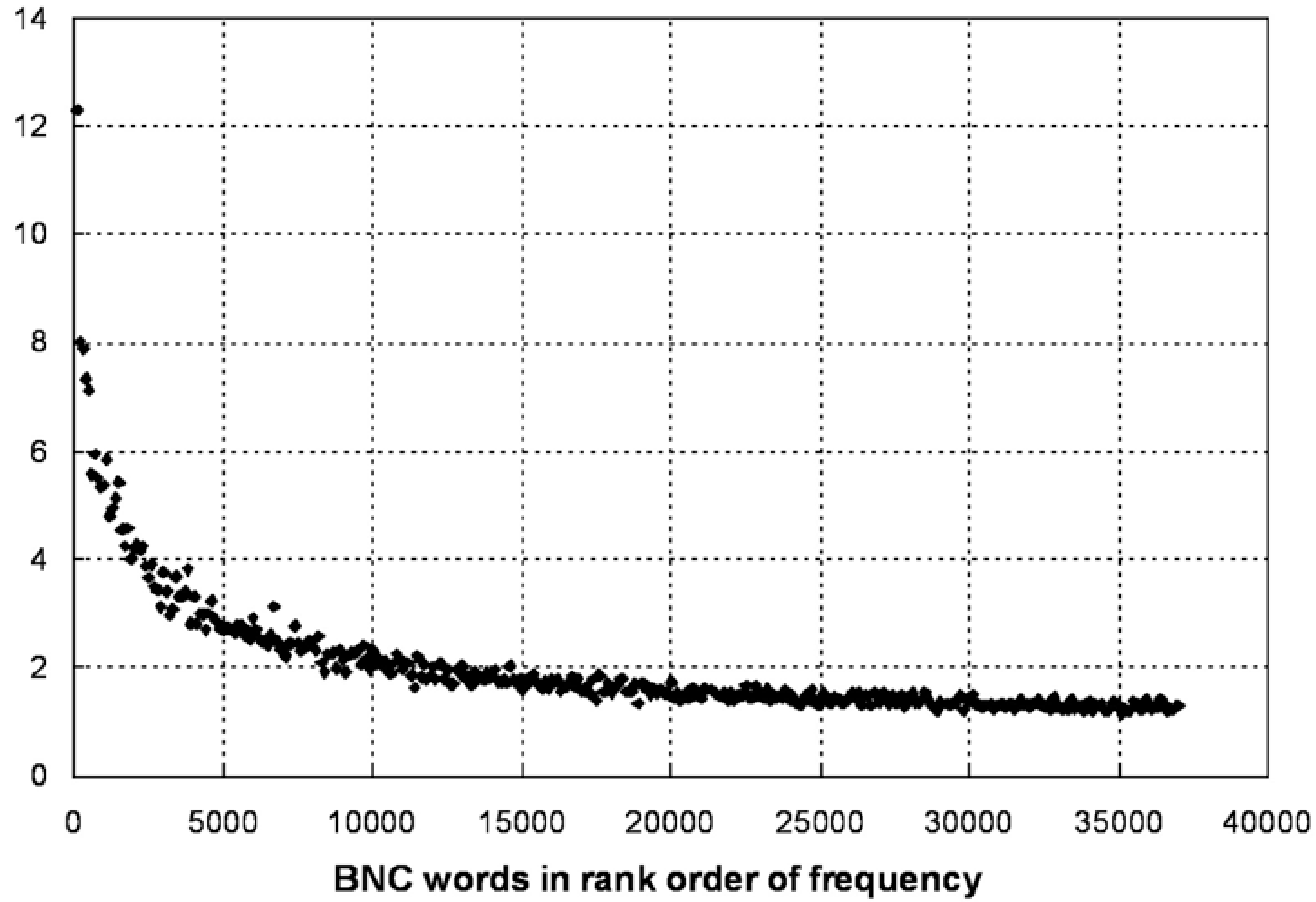
Word Sense Disambiguation

- Word sense disambiguation (WSD)
- Lexical disambiguation
- Resolving lexical ambiguity
- Lexical ambiguity resolution

Synonymy

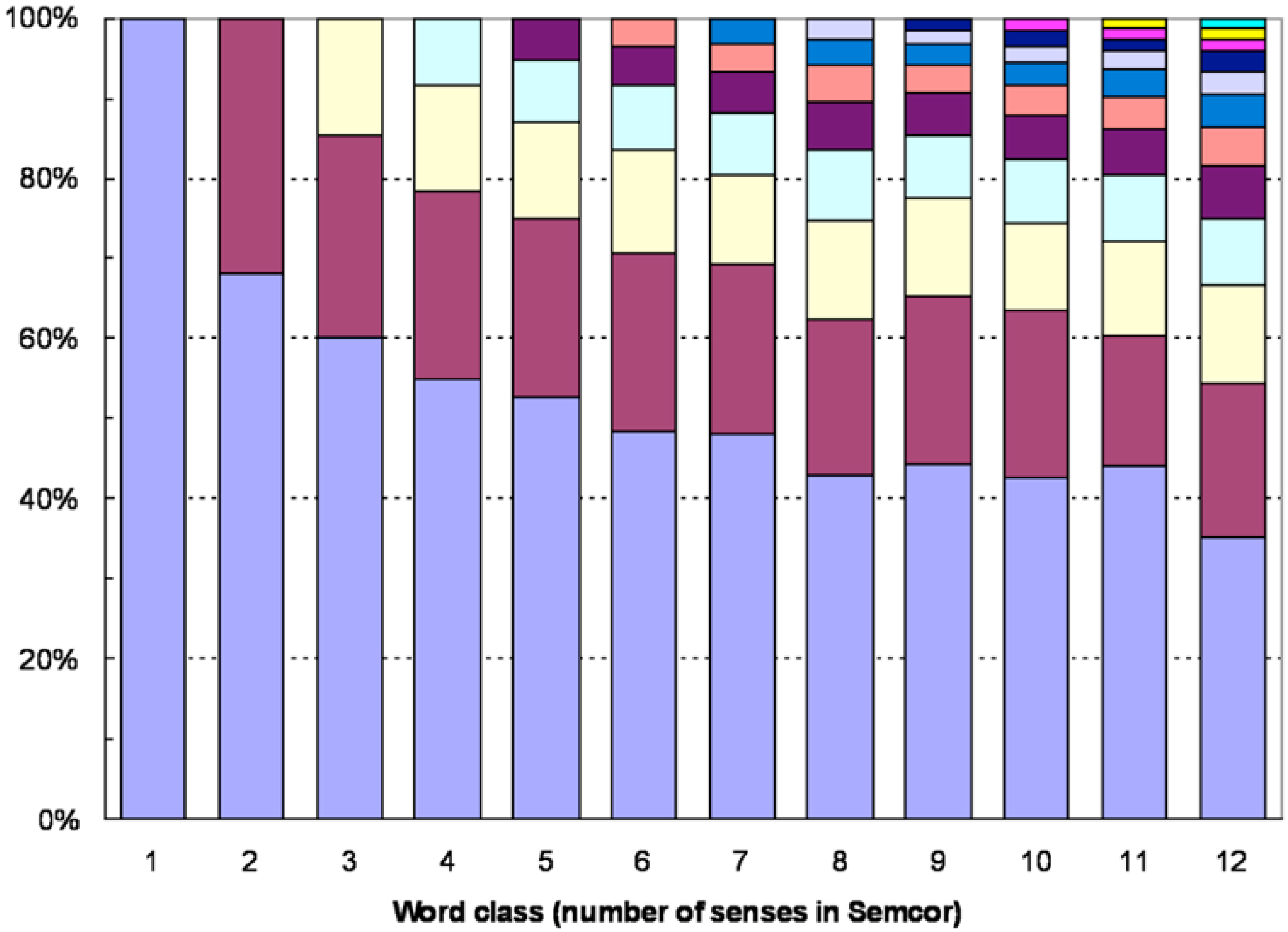
How big is the problem?

- Most words of English have only one sense
  - 62% in Longman’s Dictionary of Contemporary English
  - 79% in WordNet
- But the others tend to have several senses
  - Avg 3.83 in LDOCE
  - 2.96 in WordNet
- Ambiguous words are more frequently used
  - In British National Corpus, 84% of instances have more than one sense in WordNet
- Some senses are more frequent than others.



Words occurring in the British National Corpus are plotted on the horizontal axis in rank order by frequency in the corpus. Number of WordNet senses per word is plotted on the vertical axis. Each point represents a bin of 100 words and the average number of senses of words in the bin.

Edmonds, Philip. “Disambiguation, Lexical.” *Encyclopedia of Language and Linguistics* (second edition), Elsevier, 2006, pp 607–623.

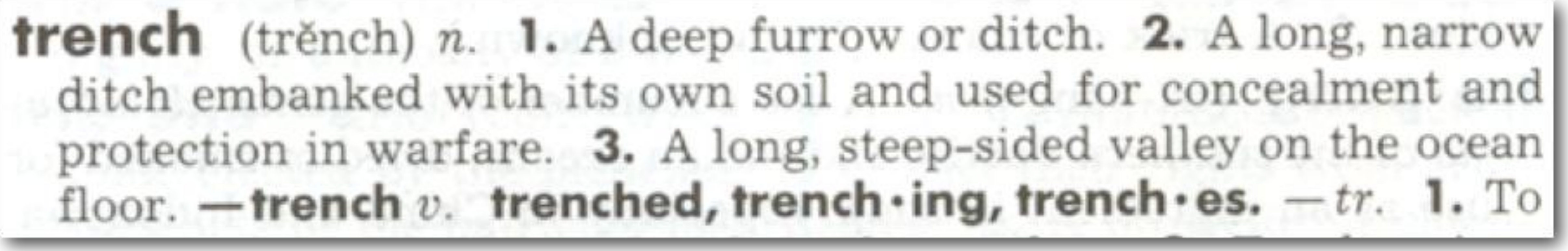


In each column, the senses are ordered by frequency, normalized per word, and averaged over all words with that number of senses.

Edmonds, Philip. “Disambiguation, Lexical.” *Encyclopedia of Language and Linguistics* (second edition), Elsevier, 2006, pp 607–623.

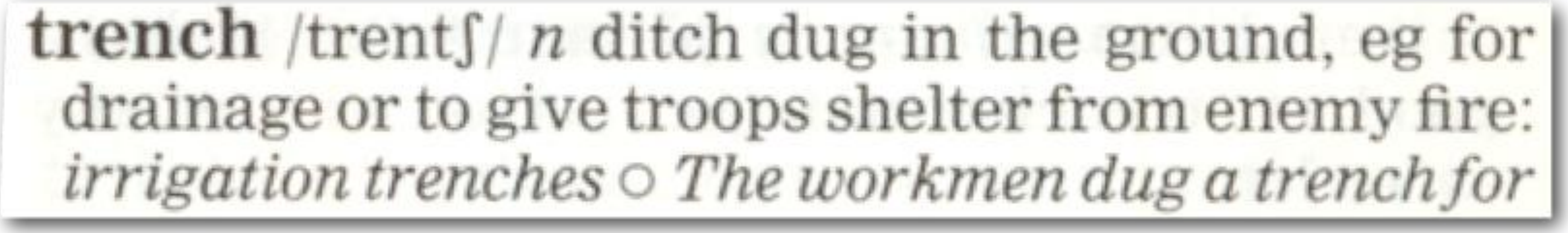
Sense inventory of a word

- Dictionaries, WordNet list senses of a word.
- Often, no agreement on proper sense-division of words.
- Don’t want sense-divisions to be too coarse-grained or too fine-grained.
    - Frequent criticism of WordNet



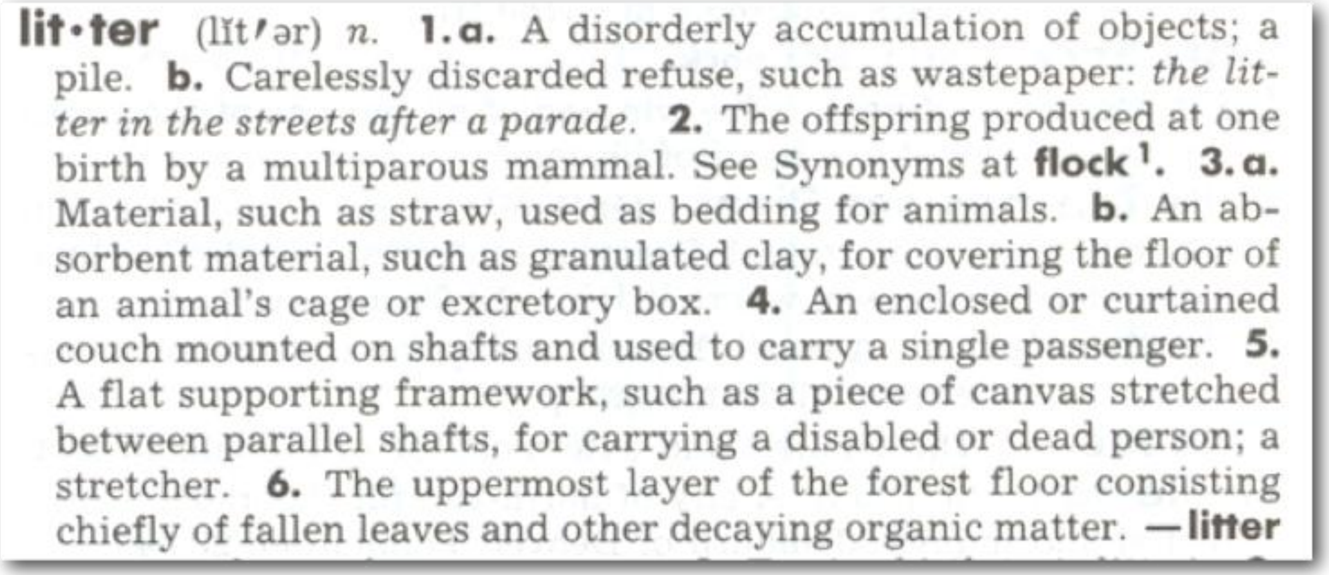
trench (trĕnch) *n.* **1.** A deep furrow or ditch. **2.** A long, narrow ditch embanked with its own soil and used for concealment and protection in warfare. **3.** A long, steep-sided valley on the ocean floor. —**trench** *v.* **trenched, trench·ing, trench·es.** —*tr.* **1.** To

The American Heritage Dictionary of the English Language (3rd edition)



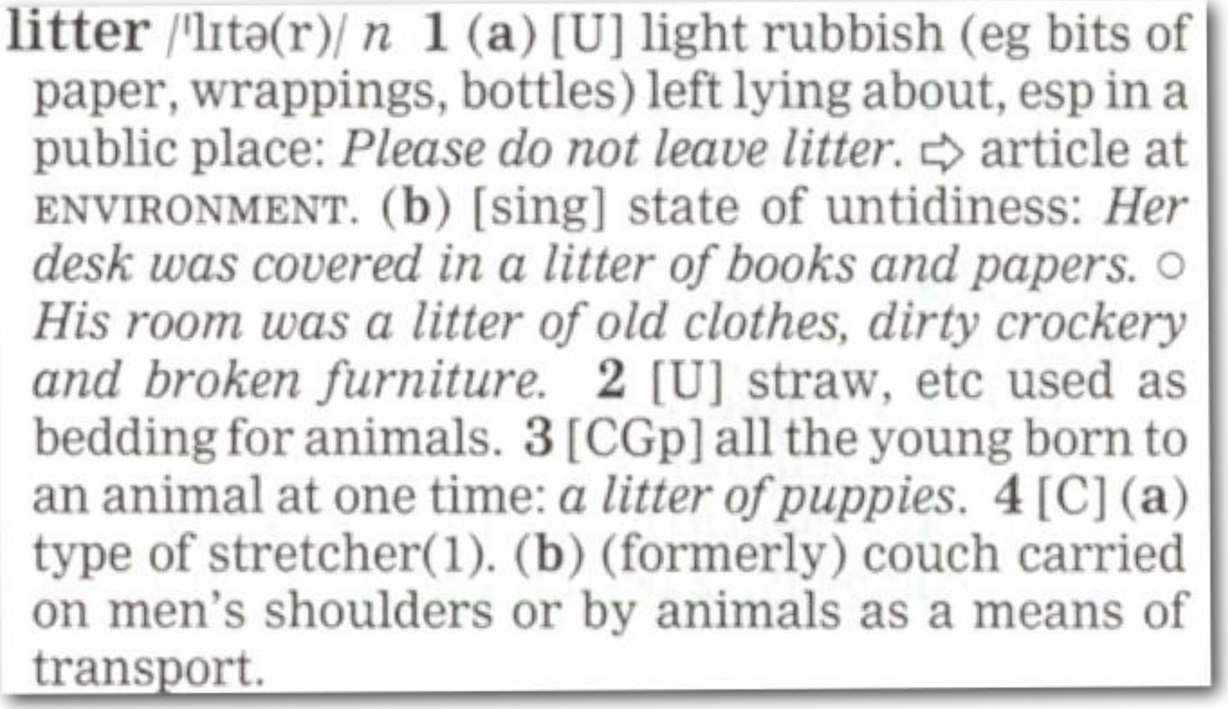
trench /trentʃ/ *n* ditch dug in the ground, eg for drainage or to give troops shelter from enemy fire: *irrigation trenches* ○ *The workmen dug a trench for*

Oxford Advanced Learner’s Dictionary (encyclopedic edition)



lit·ter (lĭt′ər) *n.* **1.a.** A disorderly accumulation of objects; a pile. **b.** Carelessly discarded refuse, such as wastepaper: *the litter in the streets after a parade.* **2.** The offspring produced at one birth by a multiparous mammal. See Synonyms at **flock**[1]. **3.a.** Material, such as straw, used as bedding for animals. **b.** An absorbent material, such as granulated clay, for covering the floor of an animal's cage or excretory box. **4.** An enclosed or curtained couch mounted on shafts and used to carry a single passenger. **5.** A flat supporting framework, such as a piece of canvas stretched between parallel shafts, for carrying a disabled or dead person; a stretcher. **6.** The uppermost layer of the forest floor consisting chiefly of fallen leaves and other decaying organic matter. **—litter**

AHDEL



litter /ˈlɪtə(r)/ *n* **1** (**a**) [U] light rubbish (eg bits of paper, wrappings, bottles) left lying about, esp in a public place: *Please do not leave litter.* ⇨ article at ENVIRONMENT. (**b**) [sing] state of untidiness: *Her desk was covered in a litter of books and papers.* ○ *His room was a litter of old clothes, dirty crockery and broken furniture.* **2** [U] straw, etc used as bedding for animals. **3** [CGp] all the young born to an animal at one time: *a litter of puppies.* **4** [C] (**a**) type of stretcher(1). (**b**) (formerly) couch carried on men's shoulders or by animals as a means of transport.

OALD

What counts as the right answer?

- Often, no agreement on which sense a given word-token is.
- Some tokens seem to have two or more senses at the same time.

Which senses are these?

- Image
  1. a picture formed in the mind;
  2. a picture formed of an object in front of a mirror or lens;
  3. the general opinion about a person, organization, etc, formed or intentionally created in people's minds;

  [and three other senses]

"… of the Garonne, which becomes an unforgettable ***image****. This is a very individual film, mannered, …"*

Example from: Kilgarriff, Adam. "Dictionary word sense distinctions: An enquiry into their nature." *Computers and the Humanities,* 26: 365–387, 1993. Definitions from *Longman Dictionary of Contemporary English*, 2nd edition, 1987.

Which senses are these?

- *distinction*
  1. the fact of being different;
  2. the quality of being unusually good; excellence.

> *"… before the war, shares with Rilke and Kafka the* ***distinction*** *of having origins which seem to escape …"*

Example from: Kilgarriff, Adam. "Dictionary word sense distinctions: An enquiry into their nature." *Computers and the Humanities,* 26: 365–387, 1993. Definitions from *Longman Dictionary of Contemporary English*, 2nd edition, 1987.

What counts as the right answer?

- Therefore, hard to get a definitive sense-tagged corpus.
- And hard to get human baseline for performance.
  - Human annotators agree about 70–95% of the time.
    [Depending on word, sense inventory, context size, discussions, etc.]

Baseline algorithms

- Assume that input is PoS-tagged.
- Obvious baseline algorithm:
  Pick most-likely sense (or pick one at random).
- Accuracy: 39–62%

Baseline algorithms

- *Simple tricks (1):*
  Notice when ambiguous word is in unambiguous fixed phrase.
  - private school, private eye.
  - All right?

But you said he was all right.
Yes. He's lost his left hand, so he's going to be "all right."

Show: Arrested Development

Baseline algorithms

- *Simple tricks (2):*
  “One sense per discourse”
  A homonymous word is rarely used in more than one sense in the same text.
    - If word occurs multiple times, …
    - Not true for **polysemy**.

- *Simple tricks (3):*
  Lesk’s algorithm (see below).

“Context”

- Meaning of word in use depends on (determined by) its context.
  - Circumstantial context.
  - Textual context.
    - Complete text.
    - Sentence, paragraph.
    - Window of n words.

“Context”

- Words of context are also ambiguous; need for mutual constraints; often ignored in practice.
- “One sense per collocation”.
- Collocation: words that *tend* to co-occur together.

Selectional preferences

- Constraints imposed by one word meaning on another—especially verbs on nouns.
  - I don't mind washing **dishes** now and then.
  - It was the most popular **dish** served in the Ladies' Grill.

- Some words select more strongly than others.
  see (weak) — drink (moderate) — elapse (strong)

see you *with eyes?*
meet?

time elapse

Limitations of selectional preferences

- Negation:

  You can't eat good intentions.
  It's nonsense to say that a book elapsed.

- Odd events:

  Los Angeles secretary Jannene Swift ***married** a 50-pound pet rock* in a formal ceremony in Lafayette Park. (Newspaper report)

- Metaphor:

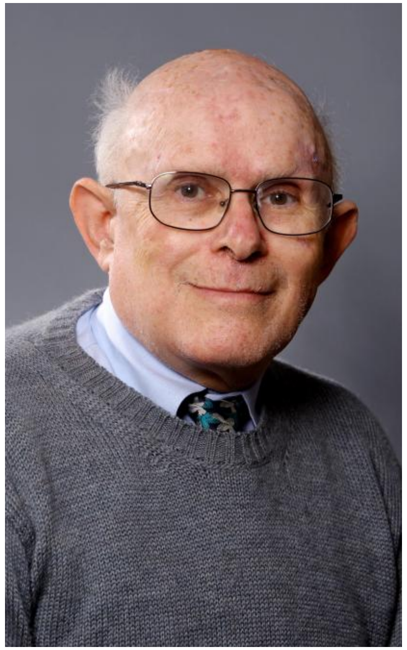
  *The issue was acute because the exiled Polish Government in London, supported in the main by Britain, was still competing with the new Lublin Government formed behind the Red Army. More time was spent in trying to **marry** these incompatibles than over any subject discussed at Yalta. … The application of these formulae could not please both sides, for they really attempted to **marry** the impossible to the inevitable.*

Limitations of selectional preferences

- In practice, attempts to induce selectional preferences or to use them have not been very successful.
    - Apply in only about 20% of cases, achieve about 50% accuracy. (Mihalcea 2006, McCarthy & Carroll 2003)
    - At best, they are a coarse filter for other methods.

Lesk’s algorithm

- Sense s_i of ambiguous word w is likely to be the intended sense if many of the words used in the dictionary definition of s_i are also used in the definitions of words in the context window.



*… the keyboard of the **terminal** was …*

terminal

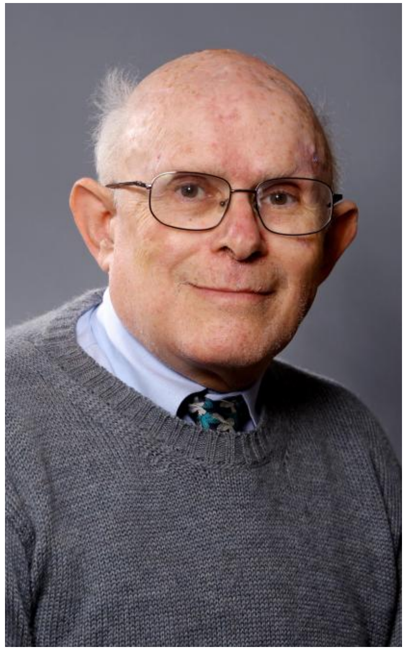
1. a point on an electrical device at which electric current enters or leaves.
2. where transport vehicles load or unload passengers or goods.
3. an input-output device providing access to a **computer**.

keyboard

1. set of keys on a piano or organ or typewriter or typesetting machine or **computer** or the like.
2. an arrangement of hooks on which keys or locks are hung.

More overlap, more likely to be the sense!

Lesk’s algorithm

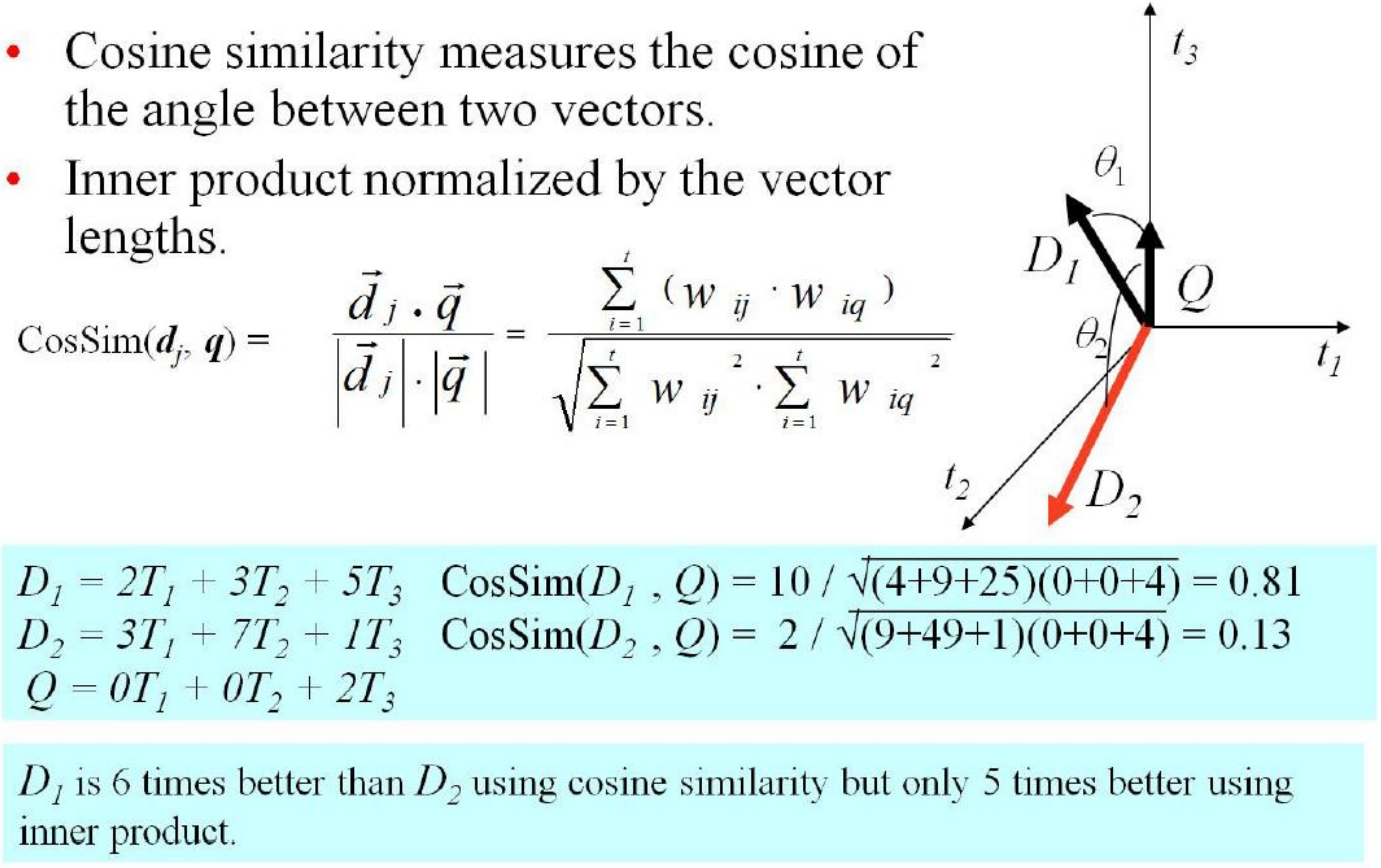


- Sense s_i of ambiguous word w is likely to be the intended sense if many of the words used in the dictionary definition of s_i are also used in the definitions of words in the context window.
- For each sense s_i of w, let D_i be the bag of words in its dictionary definition.
- **Bag of words**: unordered set of words in a string, excepting those that are very frequent (stop list).
- Let B be the bag of words of the dictionary definitions of all senses of all words $v \neq w$ in the context window of w. (Might also (or instead) include all v in B.)
- Choose the sense s_i that maximizes `overlap`(D_i, B).

Lesk’s algorithm

- Many variants of overlap score, but most common are based on **cosine similarity** of vectors that count occurrences of each word.
- **Results**:
  - Simple versions of Lesk achieve accuracy around 50–60%;
  - Lesk plus simple smarts gets to nearly 70%.
- Many variants possible on what is included in D_i and B.
  - E.g., include the examples in dictionary definitions.
  - E.g., include other manually tagged example texts.
  - PoS tags on definitions.
  - Give extra weight to infrequent words occurring in the vectors.

Cosine Similarity Score (bag of words)



- Cosine similarity measures the cosine of the angle between two vectors.
- Inner product normalized by the vector lengths.

$$\text{CosSim}(\boldsymbol{d}_j, \boldsymbol{q}) = \frac{\vec{d}_j \cdot \vec{q}}{\left|\vec{d}_j\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t}(w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^{2} \cdot \sum_{i=1}^{t} w_{iq}^{2}}}$$

t_3
θ_1
D_1
Q
θ_2
t_1
t_2
D_2

$D_1 = 2T_1 + 3T_2 + 5T_3$ CosSim(D_1 , Q) = 10 / $\sqrt{(4+9+25)(0+0+4)}$ = 0.81
$D_2 = 3T_1 + 7T_2 + 1T_3$ CosSim(D_2 , Q) = 2 / $\sqrt{(9+49+1)(0+0+4)}$ = 0.13
$Q = 0T_1 + 0T_2 + 2T_3$

D_1 is 6 times better than D_2 using cosine similarity but only 5 times better using inner product.

Cosine Similarity Score (bag of words)

- of

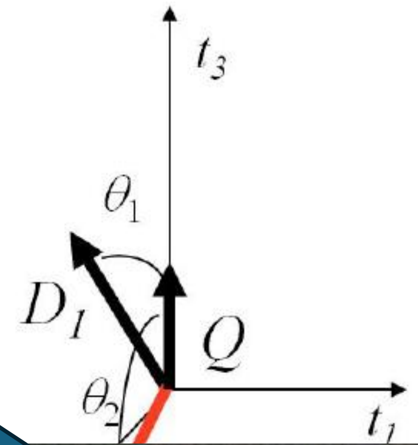
- r

C

Language model!

Word Embedding!

(more on this later)



t_2

$D_1 = 2T_1 + 3T_2 + 5T_3$ CosSim(D_1 , Q) = 10 / √(4+9+2

$D_2 = 3T_1 + 7T_2 + 1T_3$ CosSim(D_2 , Q) = 2 / √(9+49+

$Q = 0T_1 + 0T_2 + 2T_3$

D_1 is 6 times better than D_2 using cosine similarity but only
inner product.