

Week 2

NLP 與語言學的分合史

模組化，與模組化背後的語言觀

Linguistic Foundations of Large Language Models

經典教科書(推薦閱讀)，其實也反映了歷史

Jurafsky & Martin, *Speech and Language Processing* (3rd edition)

本週從一個問題出發

語言的基本單位，由誰決定？

Who gets to decide what counts as a linguistic unit?

解剖學類比

	Pipeline NLP	End-to-end LLM
Analogue discipline	classical anatomy	fMRI
What you can see	named components and interfaces	distributed activity without fixed linguistic labels
When it fails	often traceable to one stage	harder to assign to one representation or mechanism
Typical risk	cascading error	limited causal explanation

→ Interpretability is **the new neuroanatomy** (preview of Module 6)

Part 1

我們先看看 NLP 和語言學的分合史：三次擺盪

擺盪一：合（1950s-60s）

- The first wave of MT optimism — Weaver's memorandum (1949)
- Formal grammar gave symbolic NLP explicit categories, rules, and structures
- Information theory and distributional analysis supplied a second lineage: prediction from observable patterns
- Many systems shared the belief that **language could be decomposed into computable representations**
- The ALPAC report (1966) → the first winter

! 先提醒大家：語言學從來不只有一條路

Formal / generative approaches	Functional / cognitive / usage-based approaches
abstract structural system	patterns shaped by use and communicative function
discrete categories and constraints	gradient and prototype-based categories
competence distinguished from performance	frequency and experience are theoretically central
grammar licenses possible structures	grammar emerges through learned constructions

Distributional traditions add another key claim: **similar contexts support similar representations.**

擺盪二：分（1990s 統計轉向）

- Penn Treebank (1993), IBM alignment models, the Brown corpus
- The circulating version of Jelinek's line: fire a linguist, watch the accuracy climb
- In benchmark-driven NLP, one prominent role for linguists shifted from theory design to annotation and evaluation
- Many linguistic structures remained: tags, trees, alignments, senses, and semantic roles

Statistical NLP often changed the inference mechanism before it changed the representational ontology.

擺盪三：徹底的分（2013→）

- word2vec → seq2seq → Transformer → LLM
- Pretraining can learn useful representations without task-specific gold labels
- Linguists increasingly act as theory builders, evaluators, critics, and interpreters of learned systems

Question: is this separation, or a new division of explanatory labor?

但「分」不等於「沒有語言觀」

Every engineering decision is an **undeclared** linguistic commitment.

Each layer of the pipeline carries assumptions — nobody just writes them down as assumptions in the paper.

→ Part 2 takes them apart layer by layer

Part 2

NLP 管線的解剖學

Every layer encodes a linguistic assumption

- 同時可以參考對照 spaCy 的教學：<https://spacy.io/usage/spacy-101>

核心對照表（1） 單位與結構

Pipeline stage	Built-in assumption	Theoretical source	After end-to-end
tokenization	stable subunits can support later analysis	orthographic and lexical traditions	units optimize vocabulary, frequency, and sequence length
POS tagging	a finite label set captures contextual category	distributional and grammatical analysis	category information may be distributed and gradient
parsing	hierarchical relations merit explicit structures	multiple formal grammar traditions	syntactic information may be latent, without a declared tree

核心對照表（2）意義與言談

Pipeline stage	Built-in assumption	Theoretical source	After end-to-end
semantics / SRL	roles and composition merit explicit labels	formal and frame semantics	meaning is learned across contextual representations
coreference	discourse entities require explicit tracking	DRT, Centering	context supports implicit tracking, often unreliably
pragmatics	context changes inferred speaker meaning	Gricean and interactional traditions	pragmatic behavior emerges without a separate declared module

四種 modularity 常被混為一談

1. **Linguistic** modularity — distinct domains of analysis
2. **Engineering** modularity — separate components and interfaces
3. **Annotation** modularity — each label set gets its own metric
4. **Cognitive** modularity — specialized and partly encapsulated processes

Evidence for one does not establish the others.

Pipelines reflect a mixture of linguistic inheritance, available annotations, software design, and research institutions. Their shape alone cannot establish how the mind organizes language.

小結：管線是一種 **建制化** (institutional arrangement)

NLP pipeline 不只是依照語言理論自然形成的技術架構，也受到資料集、評測機制與論文發表制度的塑造。

- Shared tasks stabilize particular linguistic units
共同任務必須先規定預測單位，例如「詞」「字」「實體」或「句法依存關係」。一旦資料、工具與排行榜都建立在該單位上，後續研究便傾向繼續採用它。
- Publication practices reinforce separate processing layers
POS tagging、parsing、semantic role labeling 等任務各自可以形成資料集、模型、指標與論文。因此，語言分析的不同層次也被學術生產方式鞏固為獨立模組。
- Annotation costs constrain the granularity of categories
類別愈細，標註成本愈高、一致性愈低。標籤系統遂不只反映語言理論，也反映經費、人力與可操作性的妥協。

Chinese NLP is the clearest case of all three

Chinese NLP makes these interactions especially visible because word boundaries vary across legitimate standards.

共同任務使特定語言單位穩定下來；發表慣例強化分層處理；標註成本限制類別粒度。中文 NLP 因斷詞標準多元，而最能顯示理論、工程與研究制度如何共同塑造「語言單位」。

Part 3

中文個案：長年以來的詞彙單位之爭

「詞」不是所有語言都很明顯穩定的概念

- Ma Jianzhong, *Mashi Wentong* (1898) — helped establish *word* as a formal analytic category through engagement with European grammar
- Y. R. Chao (1968) — "word" is not a native category in Chinese; he separates
 - the sociological word
 - the grammatical word
- Xu Tongqiang, the **zi-based (字本位) position** — the basic unit should be the character

The irony: character-based neural models ended up on the zi-based side by accident.

分詞標準的不唯一性，本身就是證據

SIGHAN Bakeoff put four standards side by side: CKIP / PKU / MSRA / CTB

- One system, tested across standards → performance falls off a cliff
- Each standard's definition of "word" is internally consistent, and they are mutually incompatible

A gold standard is an **annotation convention**, not a natural kind.

- the same sentence can be segmented differently under different standards under CKIP and under jieba.

中文任務各自攜帶的預設

每一個 NLP 任務在處理中文時設定的「具體任務」，本質上都不是中立的工程工具，而是暗中背書了某種強烈的語言學哲學假設。

Task	Assumption it carries
zero-anaphora resolution	recoverable empty arguments exist at the discourse level (CTB's empty categories)
classifier prediction	noun semantics supports a classifiable ontology
N/V flexibility: 研究, 發展, 調查	with no morphology, category assignment rests on distribution alone
variant characters / encoding normalization	Unicode CJK unification = a linguistic judgment made administratively

The third row is the direct ancestor of W3's distributional semantics.

Zero-anaphora resolution (零照應還原)

- 任務：找出句子中被省略的主語或賓語（如「(他) 看了看，() 笑著說...」）。
- 假設：採用如 Penn Chinese Treebank (CTB) 的生成語法思維，預設句子底層存在一個如同英語結構般的「空範疇 (Empty Category, PRO)」，且可在篇章層次被精確還原。這把中文偏向「意合」與話題突出 (topic-prominent) 的篇章組織，強制塞入了注重句法結構完整性的西方樹狀層級。

N/V flexibility (名動兼類 / 靈活性)

- 任務：判斷「研究、發展、調查」在某個句子中究竟是名詞還是動詞。
- 假設：中文缺乏印歐語系常見的屈折變化（morphological inflections，如英語的 -ize、-ation、-ed）。因此，要將其分入詞類，唯一的判準就是它的上下文分佈（distribution）。

轉折點：斷詞還需要嗎？

- Li et al. (2019) — tests whether explicit segmentation improves deep models across Chinese NLP tasks
- Subword and byte-level tokenizers can yield units that are **neither characters nor words**

Where the authority went

linguistic analysis → annotation standards → **learned tokenization objectives**

Who decides what a linguistic unit is?

In current systems, an optimization objective and a training corpus play a major role.

Part 4

實作：同一句話，四種單位

Colab 環境準備與練習

Demo 1 — Four segmentations

```
# pip install jieba tiktoken
import jieba, tiktoken

s = "中央研究院詞庫小組的斷詞系統把這句話切開"

enc = tiktoken.get_encoding("cl100k_base")
char = list(s)
word = list(jieba.cut(s))
ids = enc.encode(s)
bpe = [enc.decode_single_token_bytes(t) for t in ids]
byte = list(s.encode("utf-8"))

for name, seq in [("char", char), ("word", word)]:
    print(f"{name:>4} ({len(seq):2d}) " + " | ".join(seq))
print(f" BPE ({len(bpe):2d}) " + " | ".join(map(repr, bpe)))
print(f"byte ({len(byte)})")
```

Demo 2 — Quantifying "objectivity"

```
def boundaries(tokens, encode=lambda x: x.encode("utf-8")):
    """Internal boundary set expressed as UTF-8 byte offsets."""
    out, i = set(), 0
    for t in tokens:
        i += len(encode(t))
        out.add(i)
    out.discard(i)          # final boundary is shared by definition
    return out

def boundary_f1(a, b):
    A, B = boundaries(a), boundaries(b)
    tp = len(A & B)
    if tp == 0:
        return 0.0
    p, r = tp / len(A), tp / len(B)
    return 2 * p * r / (p + r)

def bpe_boundaries(token_bytes):
    out, i = set(), 0
    for piece in token_bytes:
        i += len(piece)
        out.add(i)
    out.discard(i)
    return out
```

觀察到什麼

- Two accepted standards can disagree substantially, especially across registers and domains
- An LLM tokenizer optimizes a different objective, so boundary F1 may be low by design
- Downstream success therefore does not establish that token boundaries are psychologically real

A test of whether a linguistically motivated boundary helps a particular task

The uncomfortable inference problem:

can task performance serve as evidence of psychological reality?

Lab

- [Google Colab Notebook](#)

Part 5

兩種觀點的交鋒：End-to-end 的語言學意義

先說丟掉的

- Interpretable intermediate representations
- Module boundaries
- Discrete categories
- **Error attributability**

End-to-end training reduces some interface errors, but it also makes causal diagnosis harder.

但重新打開的語言理論的可能性

End-to-end learning weakened the engineering need for explicitly specified grammatical representations.

Several properties of neural language models resonate with long-standing **functional, cognitive, and usage-based** ideas:

- learning responds strongly to frequency and distribution
- categories can appear gradient and context-sensitive
- recurring form–meaning patterns resemble constructional learning
- structure can emerge from accumulated usage

Note: These affinities do not make an LLM a cognitive model or prove a linguistic theory.

The question: which tradition offers testable explanations of what the model

意外的語言學意義

Tenney et al. (2019) (BERT Rediscovered the Classical NLP Pipeline)— probing results suggest that BERT's layers roughly resemble a classical processing order:

POS → parsing → NER → SRL → coref

The declared pipeline disappeared, yet traces of its distinctions may remain in learned representations.

但是 probe 到底在測誰？

Hewitt & Liang (2019) — control tasks

- If a probe also learns random labels well, high accuracy says nothing about structure in the representation
- What matters is **selectivity**, not accuracy

Open question: can Tenney's finding and this critique both be true?

Control Task 與 Selectivity

- Hewitt & Liang (2019) 的質疑（引入 Control Task）
- Tenney 的發現 vs. Hewitt 的批判：如何同時成立？

本週結論

1. Every layer of the pipeline is an **undeclared** linguistic commitment
2. Chinese shows it plainly: linguistic units can vary across theories, standards, and learning objectives
3. LLMs reopen debate among linguistic traditions; performance alone does not adjudicate among them

Part 6

討論與作業預告

討論題

1. If character-based models reliably beat word-based ones, does that count as evidence against the psychological reality of the word? Identify exactly which step of the argument is **task-relative**.
2. Can Tenney's layer-order finding and Hewitt & Liang's control-task critique both hold?
3. LLM behavior often resembles usage-based predictions, but resemblance is not confirmation.
What evidence would connect model behavior to a linguistic theory?

參考閱讀

- Manning (2015), *Computational Linguistics and Deep Learning*
- Church (2011), *A Pendulum Swung Too Far*
- Tenney, Das & Pavlick (2019), *BERT Rediscovered the Classical NLP Pipeline*
- Li et al. (2019), *Is Word Segmentation Necessary for Deep Learning of Chinese Representations?*
- Y. R. Chao (1968), *A Grammar of Spoken Chinese* — selections on the word
- Bybee (2010), *Language, Usage and Cognition* — frequency and emergence
- Goldberg (2006), *Constructions at Work* — construction learning and generalization

Further reading: the Piantadosi (2023) / Katzir (2023) exchange; the recent Futrell & Mahowald survey

下週銜接

W3 — Tokenization and Embeddings

This week's question carries straight over:

If the unit is partly decided by an optimization objective,
what decides **meaning**?