The Computability of Language Structure
(人類語言中的數學性質)

PeterWolf
Droidtown Linguistic Tech.

Who am I?

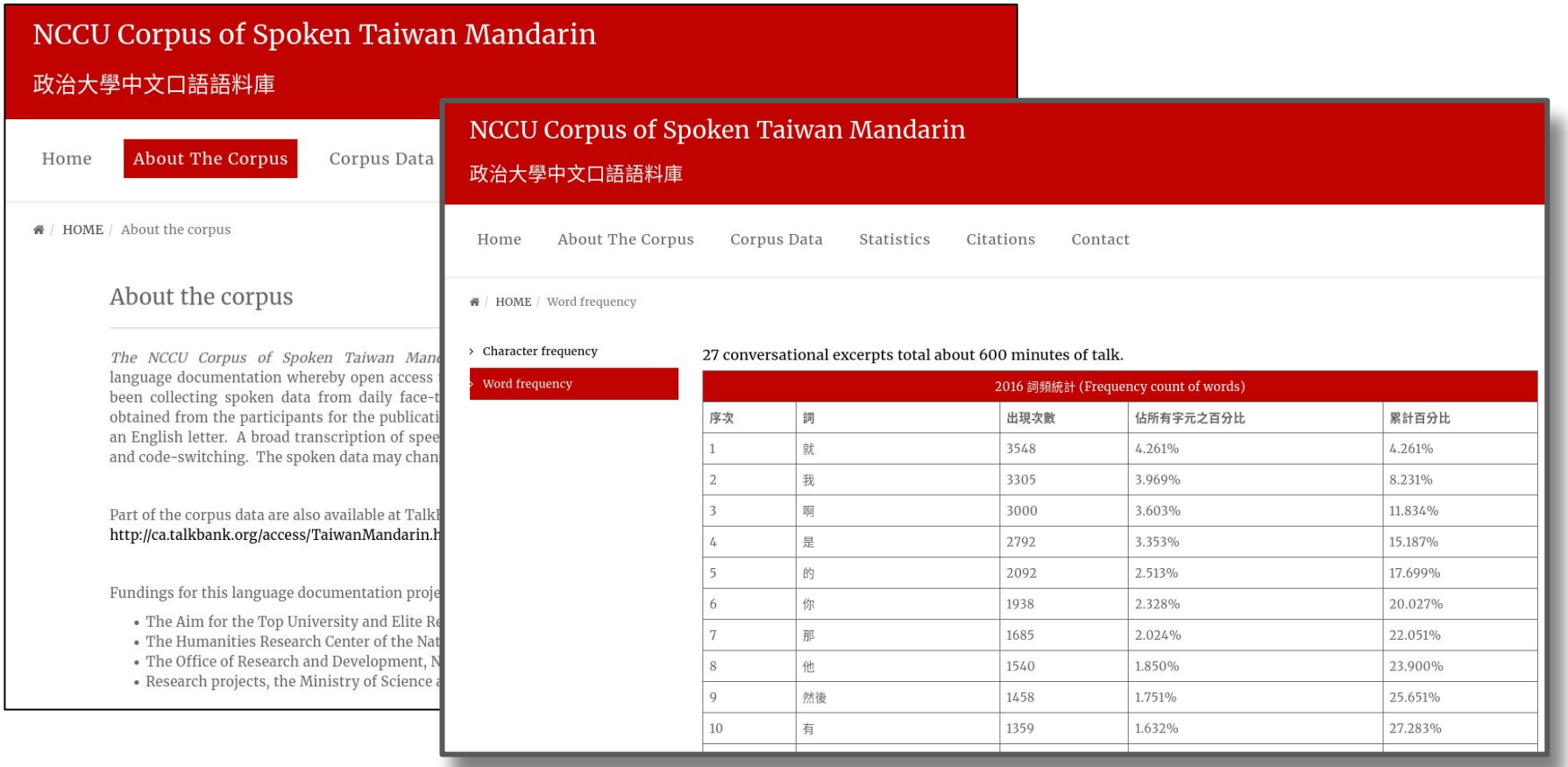
- Full-time job: Droidtown Linguistic Tech. founder & general manager.
  - Webpage: https://api.droidtown.co
  -  Discord: https://discord.gg/g5Enb5zAyK
- Inventor of linguistic technologies.
  - 3 Taiwan invention patents, and 1 US invention patent holder.
- Part-time lecturer in NYCU:
  - Linguistics and Artificial Intelligence.
  - Linguistic Technology and Complex System
  - https://github.com/PeterWolf-tw/LxTech_and_ComplexSys/

We will be talking about

MATH

today.

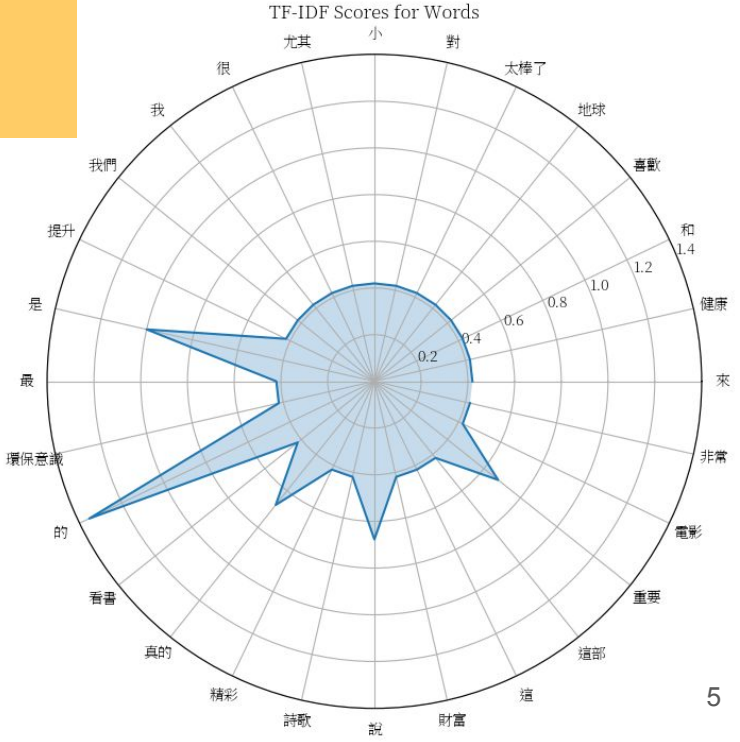
When it comes to "Math in languages"...



https://spokentaiwanmandarin.nccu.edu.tw/word-frequency.html

詞頻: TF

表示詞在文檔中出現的頻率, 就統計學而言, 只要這個詞在文本中出現越多次代表越值得關注, 因此它會具有一個重要的統計評估指標之一, 但並不是完全相信此統計方式, 看完底下的IDF就會知道為什麼。



https://vocus.cc/article/659e8a65fd897800012fefbf

https://vocus.cc/article/659e8a65fd897800012fefbf

可是為什麼很多人看成「飛越越南」？

從語言學的角度來看，一個很重要的因素發揮了影響力—**詞頻**。

首先，「飛越南部」可以拆成三個詞：

- 飛越
- 越南
- 南部

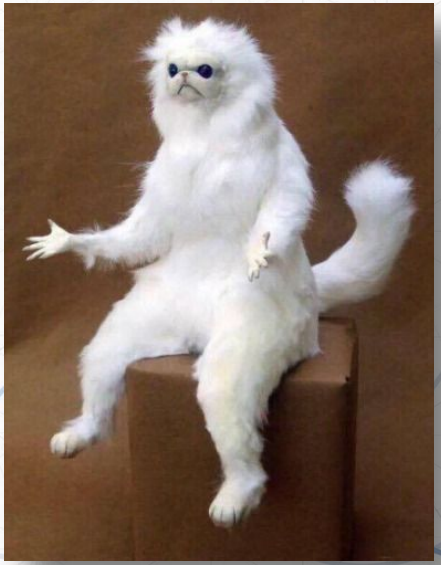
23.21 >>> 3.66，詞頻越高，人類可以越快理解這個詞，所以在「飛越」跟「越南」之間，就先取得了「越南」的資訊。

既然我們是由左至右讀，為什麼不會先抓到「飛越」這個資訊，然後順勢往下看到「南部」，組合成正確的意思？

根據中研院平衡語料庫，這三個詞的詞頻如下：

- 飛越：3.66 (每百萬字會看到 3.66 這個詞)
- 越南：23.21

Frequency! What a lovely word!



Final term paper:

請依以下兩張文字雲，從人、事、時、地、物…等角度寫出 A4 兩頁的分析報告。

背景一:這兩張文字雲是高雄市長選舉結束後，由候選人發表的演講稿製成

(但可能是不同屆哦！)

背景一:這是高雄市長選舉結束後，候選人發表的演講 (但可能是不同屆哦！)

背景二:這兩個講稿，一個是勝選人講的，一個是敗選人講的

有沒有發現，每多知道一些背景資訊，你的分析就愈具體，愈容易下筆？

文字雲？腦補＋暗示的「文本分析」

其實剛剛的文字雲是從這兩篇我自己亂寫的東西生出來的。

所以剛才都是你自己受引導以後的腦補！

一277萬一個一天一座一樣不夠不好意思之前了人口今天他鄉任但是偉大充滿內再次到剛剛勇敢包容向哪裡困難國瑜團結報告好如子弟孕育對就工作已經市民朋友希望幫幫助很很多很大忘恭喜意思愛跟感受應該打拚拜託挑戰提到時間更好替朋友期永遠熱情特別當市長疼惜相信禁結束給綠繼續美好而已能致電藍表示許多誠請變得變故鄉責任走起跟著路辜負這這塊這座這段選舉還有都長大陪伴陳其邁雖然難過韓市長願望高雄市他他作為作為其邁其邁努力努力四年四年家園家園從從或或所有所有故鄉故鄉明天明天更更最後最後生命生命相依相依這裡這裡開始開始一起一起一起也也也以以以共同共同共同加油加油加油地方地方地方市民市民市民為榮為榮為榮管管管能夠能夠能夠要要要還是還是還是城市城市城市城市感謝感謝感謝感謝未來未來未來未來為為為為人人人人人市長市長市長市長市長跟跟跟跟跟在在在在在在支持支持支持支持支持支持支持不不不不不不不不我們我們我們我們我們我們我們我們大家大家大家大家大家大家大家大家大家是是是是是是是是是我我我我我我我我我我我高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄的的的的的的的的的的的的的的的的的的的的的

一2018年一280萬一件一個一個一個一個月一定一定一定一幕一瓶一絲一起一起一起一起一遍一點上上下一心上陣下一代下來不不不不不不不停不用不用不用世界世界中中中中央也也事人人人人民今天份佛教徒你來保佑做做做做做做價值先內內內內人全全全全全力全力以赴全部兩年再出來出現分分分分族群到前前前前剛剛才力量動員包容千斤可台灣台灣台灣各位各位向向喜悅因為國民黨團結團隊團隊團隊在在在在在在城市報告報告報告外多少大家大家大家大家大家大家大家大家太多太多好好不好好好好朋友如果姊妹委員委員宣言將對手對手對手小內閣就是就是就是工作已經市民市民市民市民市民市民市長市長市長希望幫忙府廉潔廣大建立影往往往很清楚後後後得心必須志工朋友愛感謝感謝感謝感謝感謝感謝感謝感謝感謝感謝感謝感謝成長我我我我我我我我我我我我我我我我我我我我我我我我我我我們我們我們我們我們我們所以所有所有所有所有所有所有所有所有打拚打拼打造批判批判過找到找到把投票拚拚挑戰挽起提名擔心擔心支持改變政府政府政治效率敗選教育敬新明天是晚安更好更好最困難最神奇最重要最高會會會會會會會有有朝有陰朋友朋友們未來未來未來未來未來機會機會民主政治史氣水沉重沒有沒有沒有沒有沒有海深深清廉為為為為無比父特別王金平現場生活的的的的的的的的的的的的的的的的的的的的的的的的的的的的的的的的的的的的盡速直接相信相信相信相信盾盾看看到真眼礦泉祝福票數秉持立刻等待組織給給經濟經濟老兄弟肩扛肩膀能能自己與菩薩華人萬斤藍綠藍綠藍綠衝表情表情袖子要要要要要要要要親愛訊號訊號請請謝謝謝謝謝謝謝謝謝謝讓讓走走走起來跟跟跟跟跟跟跳出來輕裝逐步這個這次這裡造勢過程過程選舉選舉選舉選舉那那都都都鄉親鄉親鄉親鄉親朋友釋放重新重組重複開始關係關心院長陳其邁陳其邁陳其邁集合非常非常非常巨大非常強烈鞠躬願意願意高市府高雄市高雄市高雄市高雄市點黨部

可是為什麼很多人看成「飛越越南」？

從語言學的角度來看，一個很重要的因素發揮了影響力—**詞頻**。

首先，「飛越南部」可以拆成三個詞：

- 飛越
- 越南
- 南部

既然我們是由左至右讀，為什麼不會先抓到「飛越」這個資訊，然後順勢
部」，組合成正確的意思？

根據中研院平衡語料庫，這三個詞的詞頻如下：

- 飛越：3.66 (每百萬字會看到 3.66 這個詞)
- 越南：23.21

但為什麼不是先取得最高詞頻「南部」的資訊，然後正確理解這段話呢？

先問一個問題，「越南」跟「南部」，你覺得哪個詞的所指較精確、較具體、較容易想像？

我的答案是「越南」，如果同意的話可以繼續往下看。

語意學中有主體 (figure) 跟背景 (ground) 的概念，主體較有形、範圍較精確，背景則相反。眼睛比較容易快速看到主體，並且記住他。

既然「越南」比較具體，那麼就是因為這個原因，我們會比較先注意到詞頻高且看似為這段話主體的「越南」，而不是詞頻超高，但看似為背景的「南部」。

所以**主體-背景**的關係會凌駕於詞頻的影響嗎？答案是，我不知道，我還不知道。

computational linguistics word frequency

AI 模式 全部 圖片 影片 購物 新聞 書籍 更多 工具

ACL Anthology
https://aclanthology.org › 2020.emnlp... · 翻譯這個網頁

https://aclanthology.org/2020.emnlp-main.331/

Word Frequency Does Not Predict Grammatical ...

由 C Yu 著作 · 2020 · 被引用 18 次 — We find that across four orders of magnitude, corpus frequency is unrelated to a noun's performance on grammatical tasks.

arXiv
https://arxiv.org › cs · 翻譯這個網頁

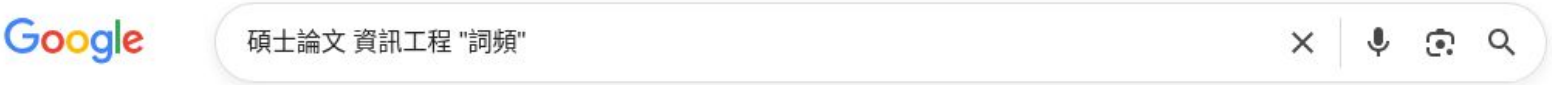
Statistical patterns of word frequency suggesting the ...

由 S Yu 著作 · 2020 · 被引用 1 次 — Abstract page for arXiv paper 2012.00187: Statistical patterns of word frequency suggesting the probabilistic nature of human languages.

Wikipedia
https://en.wikipedia.org › Zipf's_law · 翻譯這個網頁

Zipf's law

In many texts in human languages, word frequencies approximately follow a Zipf distribution with exponent · At the low-frequency end, where the rank approaches ...



Airiti Library 華藝線上圖書館
https://www.airitilibrary.com › Publication › Index

使用文字探勘實作新聞事件追蹤= News event tracking using ...

2025年8月18日 — 本論文利用R 語言建立一個新聞事件追蹤系統，透過網路爬蟲爬取新聞文章，將爬取的文章做清理，利用jieba 斷詞後，依據各文章中斷詞的結果建立詞頻矩陣，透過TF-IDF 的計算找出 ...

佛教慈濟醫療財團法人
https://dlweb01.tzuchi.com.tw › eclass › CJFD PDF

中國期刊全文資料庫中國博士學位論文全文數據庫中國優秀 ...

2025年7月13日 — 詞頻：檢索詞在相應檢索項中出現的. 頻次。詞頻為空，表示至少出現1次，. 如果為數字，例如2，則表示至少出. 現2次，以此類推. 精確：檢索結果完全等同或包. 含與檢索字 ...

中華民國圖書館學會
https://colisp.lac.org.tw › pdf › CoLISP2024 PDF

圖書資訊學術與實務研討會會議論文集

2024年12月14日 — ... 資訊技術運用於人文學研究上，產生跨領域合作. 的「數位人文」（digital humanity）研究。學者藉. 助數位工具可針對大量的文本內容進行詞頻計算，. 或是挖掘與探察某些 ...
265 頁

國立政治大學心理學系
https://psy.nccu.edu.tw › PageStaffing › Detail

蔡介立副教授 - 國立政治大學心理學系

2025年6月15日 — 第二語言經驗對閱讀第一和第二語言詞頻與詞預測力效果之調控：眼動同步記錄的大腦造影研究(延續). 蔡介立; 計畫主持人; 2020年08月~2021年07月; 國科會. 年度; 計畫名稱 ...

So, when it comes to frequency…

誰跟你說第一招先用「詞頻」

誰就是在欺負你不懂語言學！

近代語言學史

理性主義 / 形式主義

語言學是獨立的學科

語言學是理論科學
(而且很像數學)

We are here!

人類學

語言教學

歷史語言學
(主要是音韻研究)

索緒爾

傳統語言學

Chomsky 現代語言學

生成語法學派
語言可像數學一樣被計算，像科學一樣可證偽

橫空出世

布倫菲爾德學派
語言是刺激與反射

哥本哈根學派
語言是代數

布拉格學派
語言是和社會現象有關係的符號系統

計算機語言學

放棄語言學
改宗 LLM

語料庫語言學

(認知)功能學派
論文裡一定要提到 Chomsky 是錯的，語言是用來溝通的。之後寫什麼無所謂。

結構主義

行為主義 / 連結主義 / 反形式主義

If you are creating a language…

A language must have "**morphemes**" as its basic units to form words.

A language must have "**syntax**" to govern how words are put together.

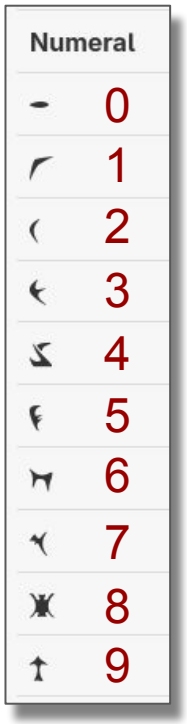
A language must have "**semantics**" to convey meaning.

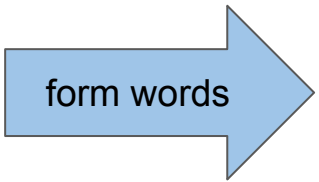
A language must have "**performance**" to deliver information (inword or outward).

Demo: Language Creator

Assuming that we have a language…part 1

- A language must have "**morphemes**" as its basic units to form words.
  - It only has 10 morphemes.





0123 => 123
246 => 246
855 => 855
56183 => 56183

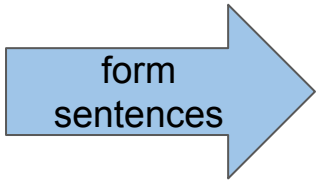
Quiz:

Is **00000** a possible word in this language?

Assuming that we have a language…part 2

- A language must have "**syntax**" to govern how words are put together.
  - It has two "functional words": +, -
  - It has embedding clauses.

0123 => 123
246 => 246
855 => 855
56183 => 56183



123 + 8550 - 24
46 + 246 - 55
85 + 8 - 5
5 + (6 - 1 + (8 + 3))

Fun facts:
In some dialects, "A + B" is noted as "A B +",
in other dialects, "A + B" is noted as "+ B A."

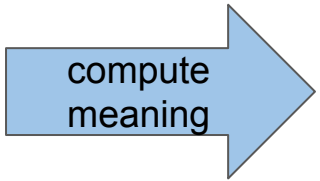
Quiz: Is this grammatical in this language?

5 + (6 - 1 + (8) + 3)

Assuming that we have a language…part 3

- A language must have "**semantics**" to compute the meaning when words are put together.
  - It has two "functional words": +, -, x
  - It has embedding clauses.

123 + 8550 - 24
46 + 246 - 55
85 + 8 - 5
5 + (6 - 1 + (8 + 3))



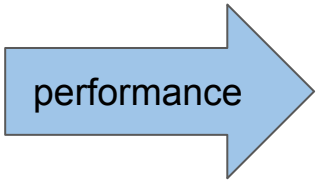
123 + 8550 - 24 = 8649
46 + 246 - 55 = 237
85 + 8 - 5 = 88
5 + (6 - 1 x (8 + 3)) = 0

Quiz: Do you think we come up with the number after "**=**" by some computational process or some probability model?

Assuming that we have a language…part 4

- A language must have "**performance**" to deliver information (inward or outward).

123 + 8550 - 24 = 8649
46 + 246 - 55 = 237
85 + 8 - 5 = 88
5 + (6 - 1 x (8 + 3)) = 0



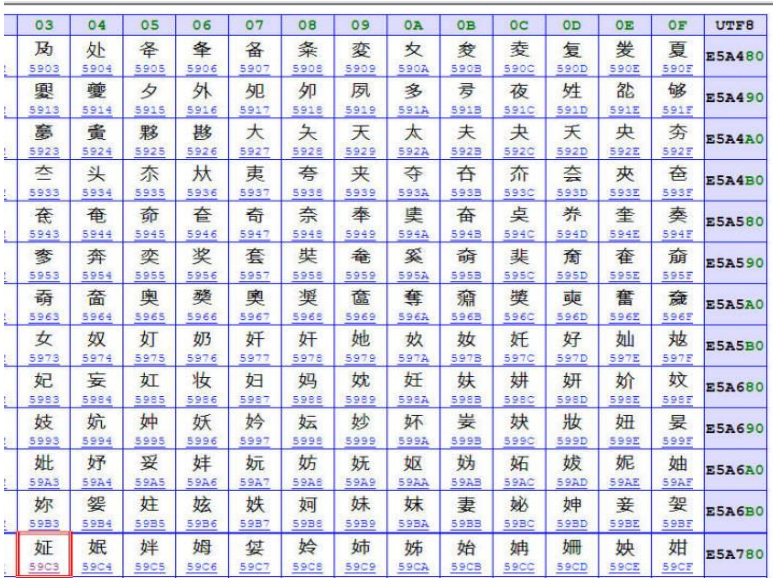
123 + 8550 - 24 = 8649
-> 8 thousand 6 hundred and 4ty-9
-> 8ty-6 4ty-9
-> 8 6 4 9

Quiz: Do you agree with the idea that

"**MATH is just another LANGUAGE?**"

We do have a language…part 1

- A language has "morphemes" as its basic units to form words.



03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	UTF8
及 5903	处 5904	冬 5905	务 5906	备 5907	条 5908	变 5909	女 590A	复 590B	变 590C	复 590D	发 590E	夏 590F	E5A480
娶 5913	娶 5914	夕 5915	外 5916	夘 5917	卯 5918	夙 5919	多 591A	孛 591B	夜 591C	姓 591D	够 591E	够 591F	E5A490
豢 5923	夤 5924	夥 5925	趁 5926	大 5927	头 5928	天 5929	太 592A	夫 592B	央 592C	天 592D	央 592E	夯 592F	E5A4A0
亖 5933	头 5934	亦 5935	从 5936	夷 5937	夸 5938	夹 5939	夺 593A	奋 593B	亦 593C	奁 593D	夾 593E	奋 593F	E5A4B0
奄 5943	奄 5944	奋 5945	奋 5946	奇 5947	奈 5948	奉 5949	奎 594A	奋 594B	卖 594C	券 594D	奎 594E	奏 594F	E5A580
奢 5953	奔 5954	奕 5955	奖 5956	套 5957	奘 5958	奄 5959	奚 595A	奋 595B	奘 595C	奋 595D	奄 595E	奋 595F	E5A590
奋 5963	奋 5964	奥 5965	奥 5966	奥 5967	奖 5968	奩 5969	奪 596A	奋 596B	奖 596C	奥 596D	奮 596E	奋 596F	E5A5A0
女 5973	奴 5974	奶 5975	奶 5976	奸 5977	奸 5978	她 5979	奴 597A	奴 597B	奼 597C	好 597D	奾 597E	奴 597F	E5A5B0
妃 5983	妄 5984	妊 5985	妆 5986	妇 5987	妈 5988	妩 5989	妊 598A	妹 598B	妍 598C	妍 598D	妁 598E	妓 598F	E5A680
妓 5993	妩 5994	妯 5995	妖 5996	妗 5997	妘 5998	妙 5999	妒 599A	妾 599B	妹 599C	妝 599D	妞 599E	妥 599F	E5A690
妣 59A3	妤 59A4	妥 59A5	妦 59A6	妩 59A7	妨 59A8	妩 59A9	妪 59AA	妫 59AB	妬 59AC	妓 59AD	妮 59AE	妯 59AF	E5A6A0
妳 59B3	娶 59B4	妊 59B5	妶 59B6	妶 59B7	妸 59B8	妹 59B9	妹 59BA	妻 59BB	妣 59BC	妽 59BD	妾 59BE	娶 59BF	E5A6B0
妊 59C3	妮 59C4	妦 59C5	姆 59C6	娶 59C7	姈 59C8	姊 59C9	姊 59CA	始 59CB	姊 59CC	姗 59CD	姎 59CE	姐 59CF	E5A780

form words

夯、奄 => ??
大、夫 => 大夫
夏、天 => 夏天
大、姊、姊 => 大姊姊

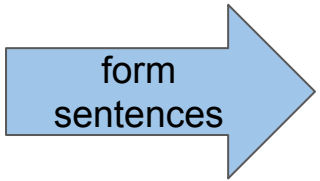
Quiz:

Why "夯奄" is unlikely to be a possible word in this language?

Assuming that we have a language…part 2

- A language must have "syntax" to govern how words are put together.
  - It has two "functional words": +, -
  - It has embedding clauses.

夯、奄 => ??
大、夫 => 大夫
夏、天 => 夏天
大、姊、姊 => 大姊姊



大姊姊 在 夏天 當上 大夫

大姊姊 聽到 (妹妹 叫 (弟弟 去 游泳))

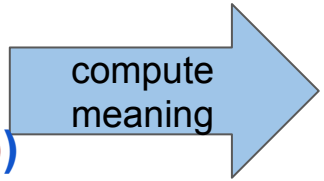
Fun facts:
In some languages, "姐姐 當上 大夫" is noted as "姐姐 大夫 當上",
in other languages, "姐姐 當上 大夫" is noted as "當上 大夫 姐姐"

Assuming that we have a language…part 3

- A language must have "semantics" to compute the meaning when words are put together.
  - It has two "functional words": +, -, x
  - It has embedding clauses.

大姊姊 在 夏天 當上 大夫

大姊姊 聽到 (妹妹 叫 (弟弟 去 游泳))

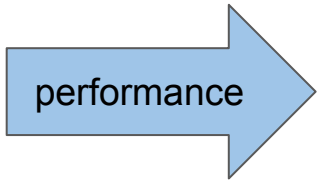


[[大姊姊在夏天當上大夫]]
= the condition in which there is a 姊姊 and a 大夫 in between a 當上-relation holds in the time when it is 夏天

Assuming that we have this language…part 4

- What are inward/outward "performance"?

[[大姊姊在夏天當上大夫]]
= the condition in which there is a 姊姊 and a 大夫 in between a 當 上-relation holds in the time when it is 夏天



You can either "think of" it in your mind, "speak it out" with your mouth or sign it with a sign language.

Fun facts:
Since "speak it out/sign it" are only part of the possible performance options, communication is not the core part of language system.

Language = Math

The core of a language system is a set of intertwined computational processes describing…

- How words are formed with morphemes
- How sentences are formed with words?
- How meanings are computed with sentences?
- How the results are performed?
- **Language is a math system with more than the numbers from 0 to 9.**

Why did computational linguistics choose frequency as the tool?

- With the birth of computers in the **1940s**, computational linguistics started.
- Chomsky's first paper came out at late **1950s**.
- When computational linguists started to investigate languages, the only tool in hand was…basically counting words as "**word frequency**."
- Many students nowadays confuse the order of the years and think that "Chomsky is doing traditional linguistics" and "computational linguists are doing modern linguistics." Na…Computational linguistics came first and the approaches are still pretty much the same as their ancestors in the 40s.

Do frequency models reflect human minds?

Large language models are said to be a compression model of the world. (ref.)

- **Language Modeling Is Compression**

  Gregoire Deletang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, Joel Veness

  Published: 16 Jan 2024, Last Modified: 15 Mar 2024 ICLR 2024 poster Everyone Revisions BibTeX

  Code Of Ethics: I acknowledge that I and all co-authors of this work have read and commit to adhering to the ICLR Code of Ethics.
  Keywords: lossless compression, arithmetic coding, language models, scaling laws, in-context learning
  Submission Guidelines: I certify that this submission complies with the submission instructions as described on https://iclr.cc/Conferences/2024/AuthorGuide.

- NOTONLY AI
  https://notonly-ai.com › blog-detail › open-ai-首席技術...

  Open AI 首席技術官Ilya Sutskever-GPT是壓縮是壓縮全世界 ...

  2023年5月10日 — 神經網絡與GPT"有一種誤解，認為ChatGPT是一個大型語言模型，但有一個圍繞它的系統，"黃仁勳說。 Sutskever表示，OpenAI使用兩個級別的訓練 ...

- THE NEW YORKER 100

  ANNALS OF ARTIFICIAL INTELLIGENCE

  CHATGPT IS A BLURRY JPEG OF THE WEB

  OpenAI's chatbot offers paraphrases, whereas Google offers quotes. Which do we prefer?

  By Ted Chiang
  February 9, 2023

Two ways of compression:

- Frequency-based compression => zip up the "frequently-seen" elements.

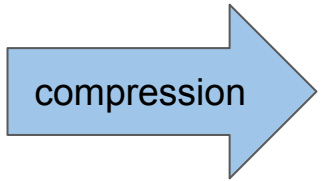
List of the most popular given names in South Korea

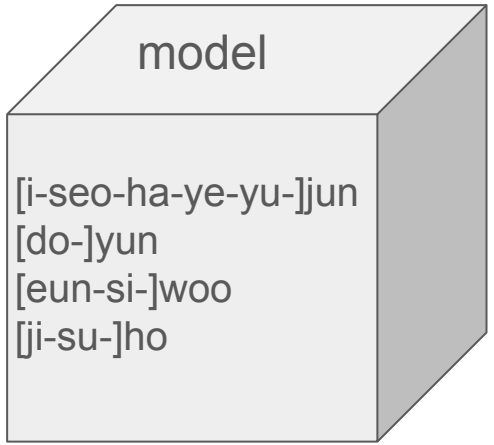
2021 [edit]

Boys[2]

Common spelling	Hangul	RR	MR	Count
I-jun	이준	Ijun	Ijun	2,833
Seo-jun	서준	Seojun	Sŏjun	2,396
Ha-jun	하준	Hajun	Hajun	2,227
Do-yun	도윤	Doyun	Toyun	2,199
Eun-woo	은우	Eunu	Ŭnu	1,931
Si-woo	시우	Siu	Siu	1,831
Ji-ho	지호	Jiho	Chiho	1,606
Ye-jun	예준	Yejun	Yejun	1,455
Yu-jun	유준	Yujun	Yujun	1,380
Su-ho	수호	Suho	Suho	1,360

69 characters in length

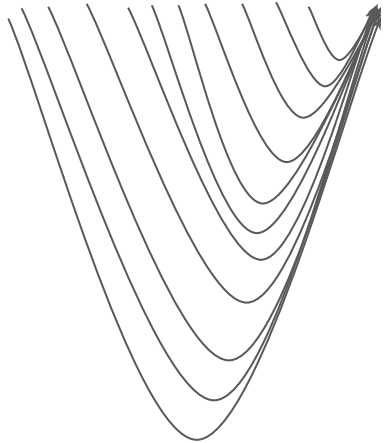




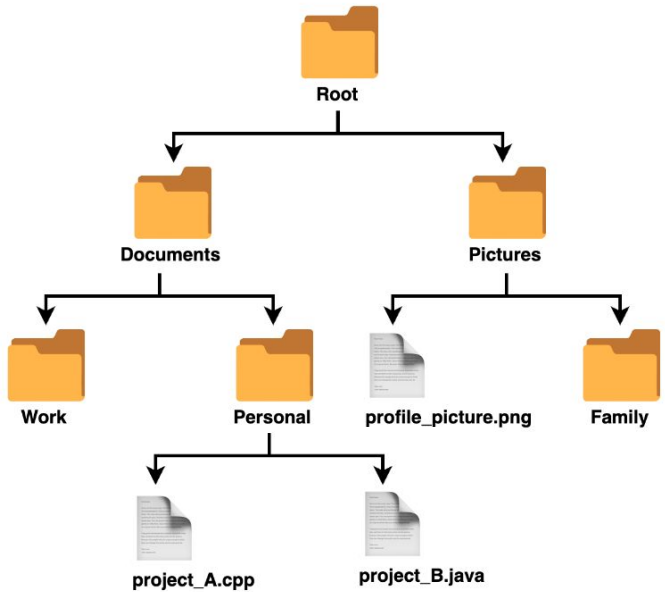
50 characters in length

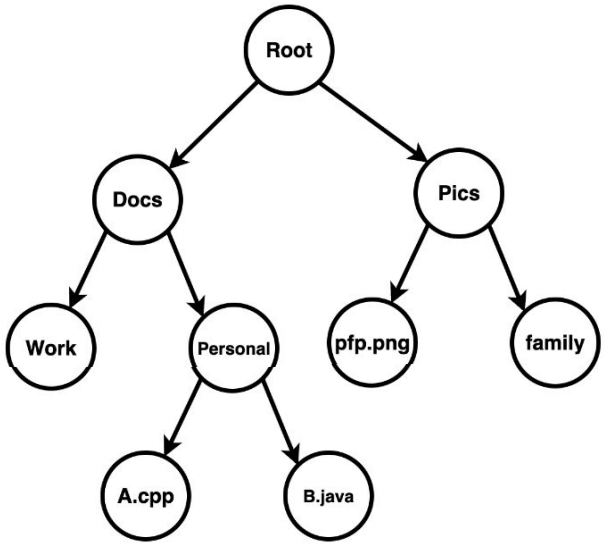
Why frequency model works, and why it is **NOT** how humans do.

- To be able to "compute" a frequency means you need "a lot of samples."
  - Human children don't have that.
- A frequency model "predicts the next token" **by probability** and that is why LLMs are bad at doing math.
  - **123 + 8550 - 24 = 8649**

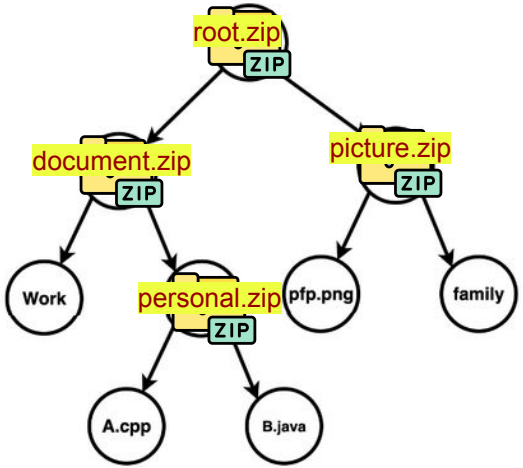


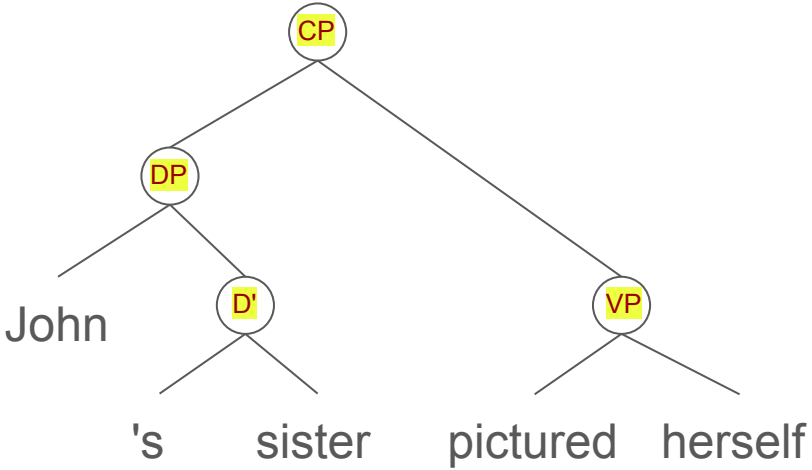
Another kind of compression: Binary-Branching Syntax Tree



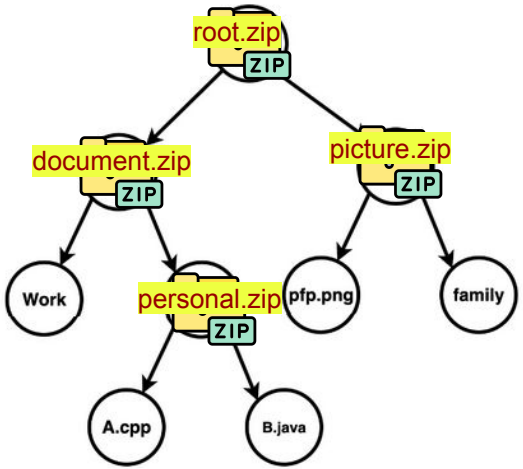


Another kind of compression: Binary-Branching Syntax Tree





A compressed object must be decompressable!



root.zip => **{document.zip, picture.zip}**

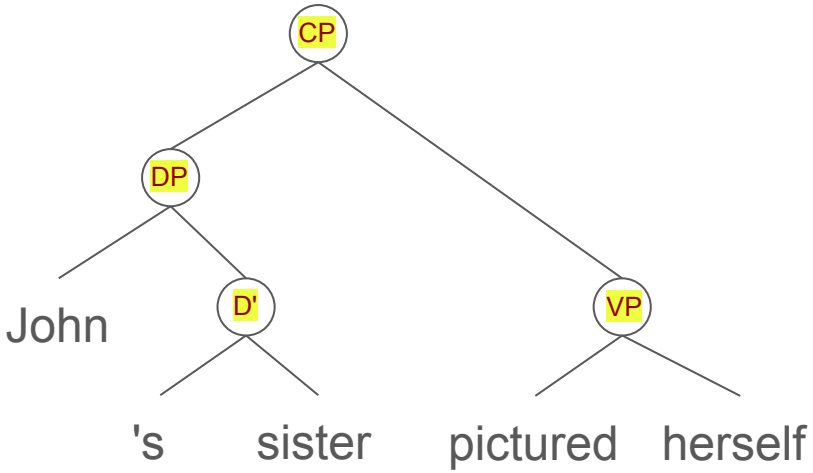
document.zip => **{work, personal.zip}**

personal.zip => **{A.cpp, B.java}**

picture.zip => **{pfp.png, family}**

- To make sure what are decompressed is the same as what were compressed before, an **ALGORITHM** that can guarantee the output instead of a **probability model** that has no guarantees to the output must be implemented.

A Binary-Branching Syntax Tree is **the Algorithm**!



CP => **{DP*, VP}**

DP => **{John, D'}**

D' => **{'s, sister}**

VP => **{pictured, herself}**

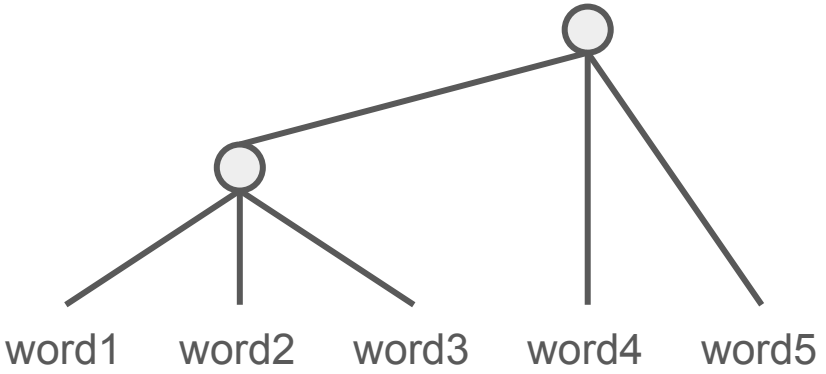
- Don't get misled! This is not the phrase-structure rules from the last century in many aspects:
  - It's **binary**: memory cost at every level is fixed.
  - It's **universal**: the compression/decompression algorithm applies to all human languages.
  - It's **corpus-free**: no data model is needed for training.
  - It's **efficient**: (to be discussed later)
  - *one more thing, it is only taught in NYCU in Taiwan.*

***Due to space limitations, this is a simplified expression. A CP should consist of {C, TP}. Audiences are encouraged to take a syntax course at NYCU for more details.**

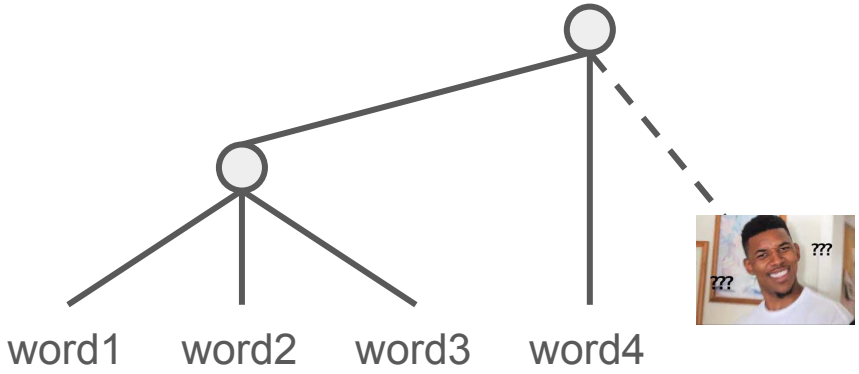
Why not ternary (or more)?

Because you cannot know how many words will be used in a sentence before you perceive it.

- Assuming you have a ternary tree in your head…



Ternary tree works perfect with a 5-word sentence.

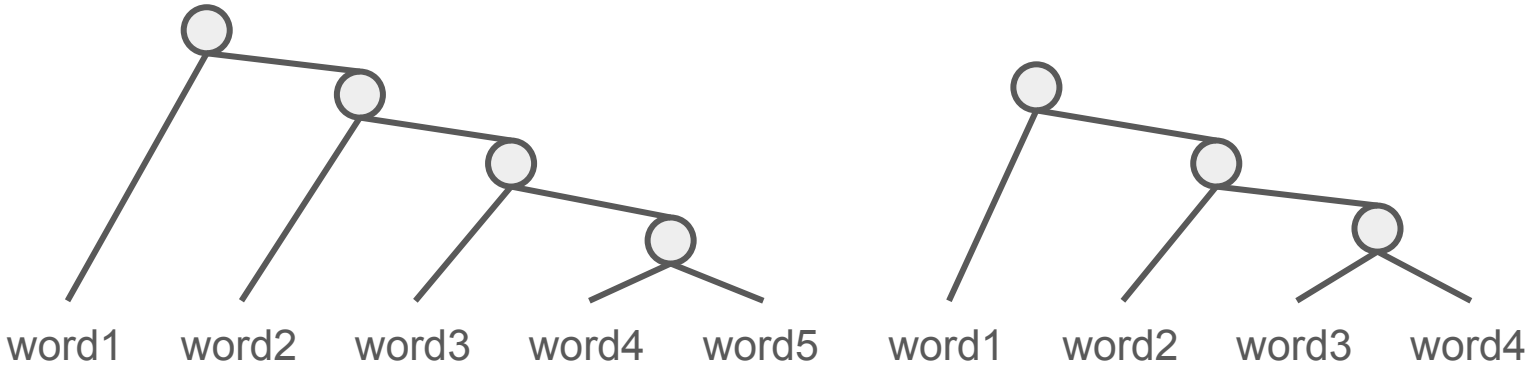


But in a 4-word sentence…

On the other hand, a binary-branching structure…

No matter how many words will be used in a sentence, a binary-branching structure can handle it.

- Assuming you have a ternary tree in your head…



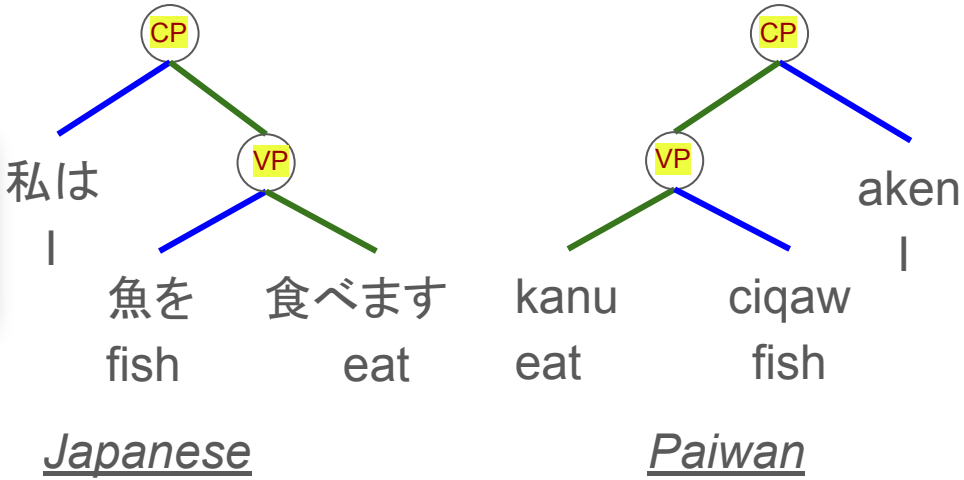
Why binary?

- When designing an algorithm, fixed data length is generally faster than non-fixed data length in computing because it simplifies and speeds up operations like memory access and instruction decoding.
- A fixed-binary-branching tree can handle sentences of different lengths.
- Fixed-ternary (or N-ary where N>2) branching tree can not handle sentences of different lengths.
- Given the fact that human languages use sentences of different length, binary-branching tree is the only viable structure.

Why "Being Universal" is important?

- We are all Homo sapiens; we are the same species.
- Universality guarantees the potentiality of a human to acquire/learn **ANY** human language in the world.

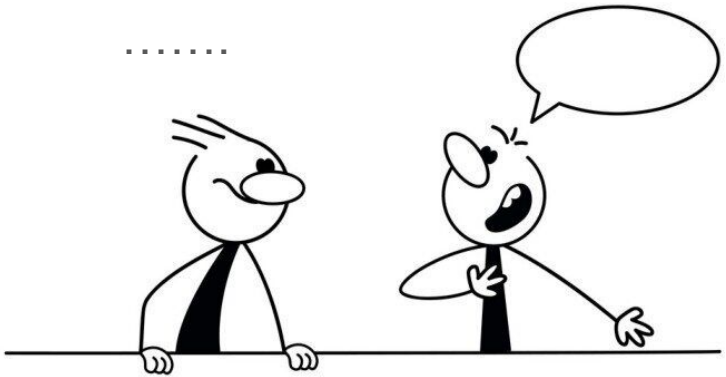
Fun facts:
In some dialects, "A + B" is noted as "A B +",
in other dialects, "A + B" is noted as "+ B A."



Why do we say binary-branching structure is innate?

When a human starts to use language to communicate with another human being, the first assumption would be that "**WE MUST SHARE THE SAME DECODING/ENCODING SYSTEM**."

We don't need to have the same experience in order to have the same language model in mind.



LLM Transfer learning, how promising it is?

arXiv:2501.11496v1 [cs.CL] 20 Jan 2025

Generative AI and Large Language Models in Language Preservation: Opportunities and Challenges

Vincent Koc

Vincent Koc is with Hyperthink, Sydney, Australia (hyperthink.com.au). Contact: vincentkoc@ieee.org.

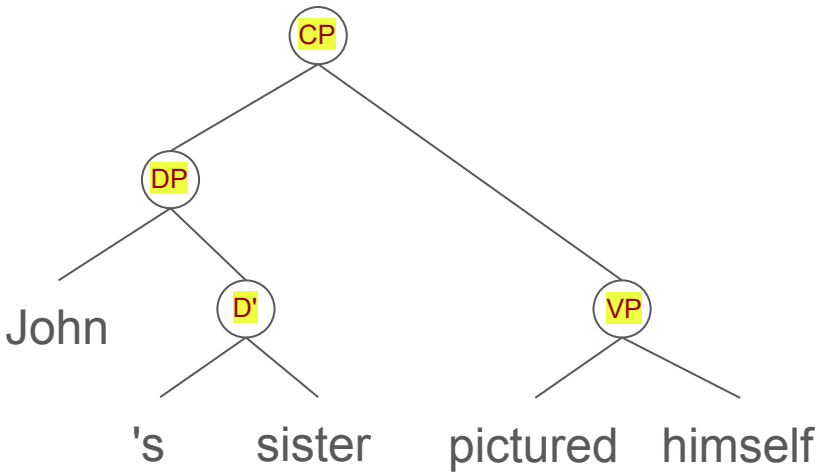
https://arxiv.org/html/2501.11496v1

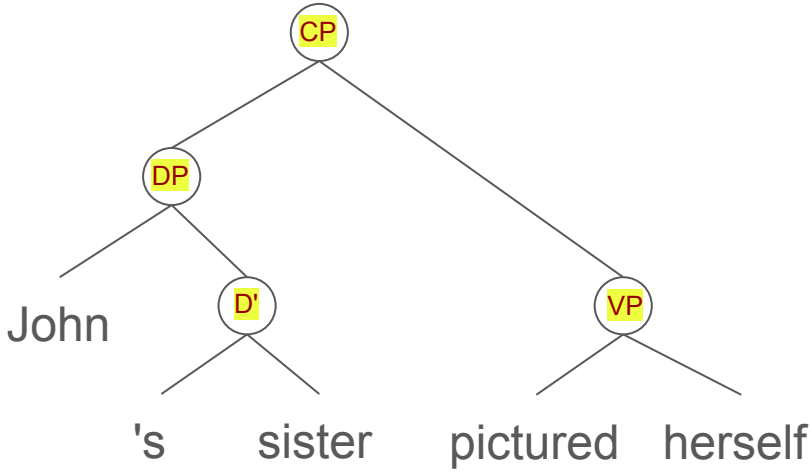
- LLM stands for "LARGE language model", it requires a "LARGE" amount of training data (and electricity) . Endangered languages usually don't have this luxury.
- LLM/AI is not a universal approach.
- If it digitizes/generates less-accurate outputs, then just please don't start. Once wrong corpus is established/generated, it will be there forever.

IV-B *Technical Limitations*

Training large-scale language models requires significant computational resources, including specialized hardware and large amounts of electricity. Access to the right infrastructure can be an obstacle, especially for endangered languages spoken in low-resource regions. Additionally, AI models often struggle with complex grammar, non-standard spellings, or extensive lexical borrowing from dominant regional languages, leading to less-accurate outputs.

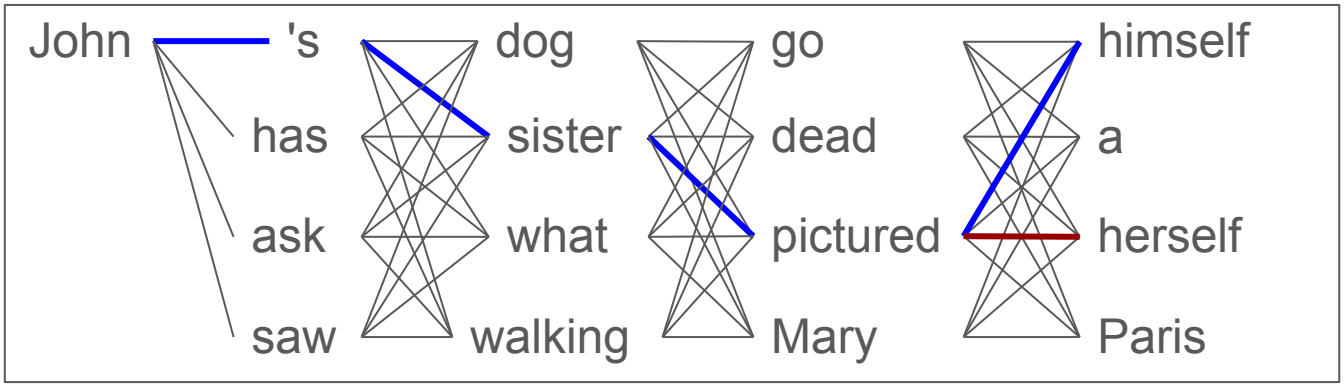
Efficiency: Which one is grammatical?





Two ways to determine whether "himself" is grammatical here

- Let's train a language model to predict whether it is more frequently to see "himself" or "herself" here.

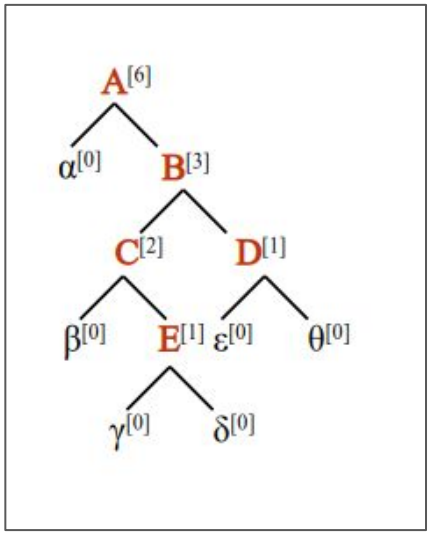


- Why is "**herself**" grammatical but "**himself**" ungrammatical here?
  - Because the probability of "herself" is higher than "himself." (←That's description, not explanation.)
- Why is "**a**" ungrammatical here? Don't you think "**pictured a …** " would be more frequent than "**pictured herself**" in any "well-balanced" corpus?

Another kind of computation: C-Command

Definition of C-Command:

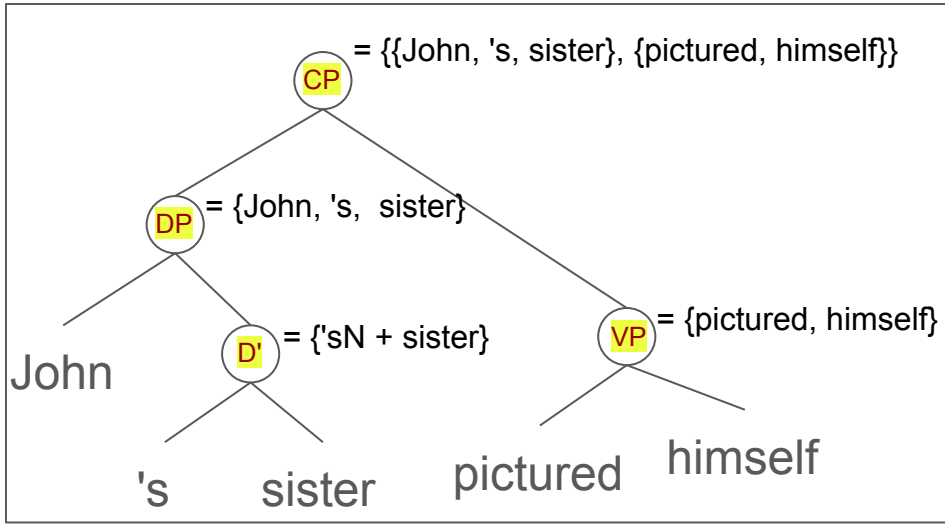
- Node α c-commands a node β iff
- Neither α nor β dominates the other, and
- Every branching node dominating α also dominates β

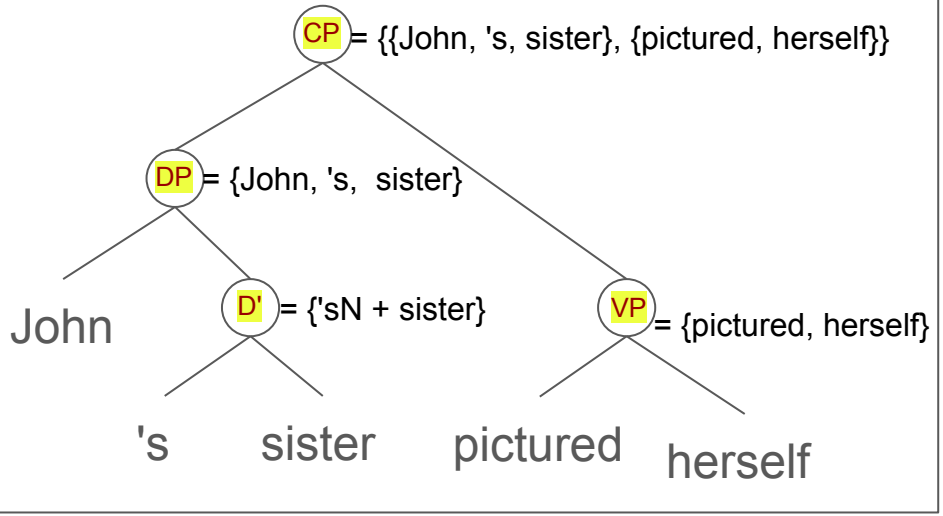


Red texts are compression nodes.

Computation with Algorithm:

- Why is "**herself**" grammatical but "**himself**" ungrammatical here?
  - Referent of **pronominal expressions** needs to be resolved.
  - The compression node containing "**herself**" is compressed with **{John, 's, sister}** but not John.
  - The compression node containing "**himself**" is compressed with **{John, 's, sister}** but not John.





- Is it just some gender-agreement issue? No, it's the structure issue! Consider this:
  - Mary's sister took a selfie by the tower. (Who is in the selfie? Mary or the sister?)

Comparison:

- Frequency-based Approach
  a. EXTREMELY large training data is required.
  b. MASSIVE electricity is required for computing.
  c. DIFFICULT to take care of endangered languages.
  d. The output is NOT EXPLAINABLE.
  e. Needs a lot of $$ investment, like…

  A LOT!

- Algorithm-based Approach
  a. Minimum data is needed (at least one instance).
  b. My old 2017 macbook air can do the computation.
  c. See (a). above.
  d. The algorithmic logic is transparent and explainable. Explainable AI means Responsible AI.
  e. Need linguists who know the structures of languages to do the programming.

Takeaways

1. **Language as an Algorithmic System**
    - Language operates like mathematics—through computational processes, not probability
2. **Binary-Branching Syntax Trees as Compression**
    - Human language structure uses **binary-branching trees** as a universal compression/decompression algorithm
    - This is the only structure that can handle variable-length sentences efficiently with fixed memory costs
    - Universal across all human languages (e.g., Japanese, Paiwan, English all use the same underlying structure)
3. **Practical Implications**
    - The C-command algorithm explains grammaticality (e.g., "herself" vs "himself") through structural relationships, not frequency
    - LLMs are not suitable for endangered language preservation—they risk generating inaccurate corpus that persists forever
    - True language digitization requires understanding linguistic structure, not just statistical patterns.

Discussions