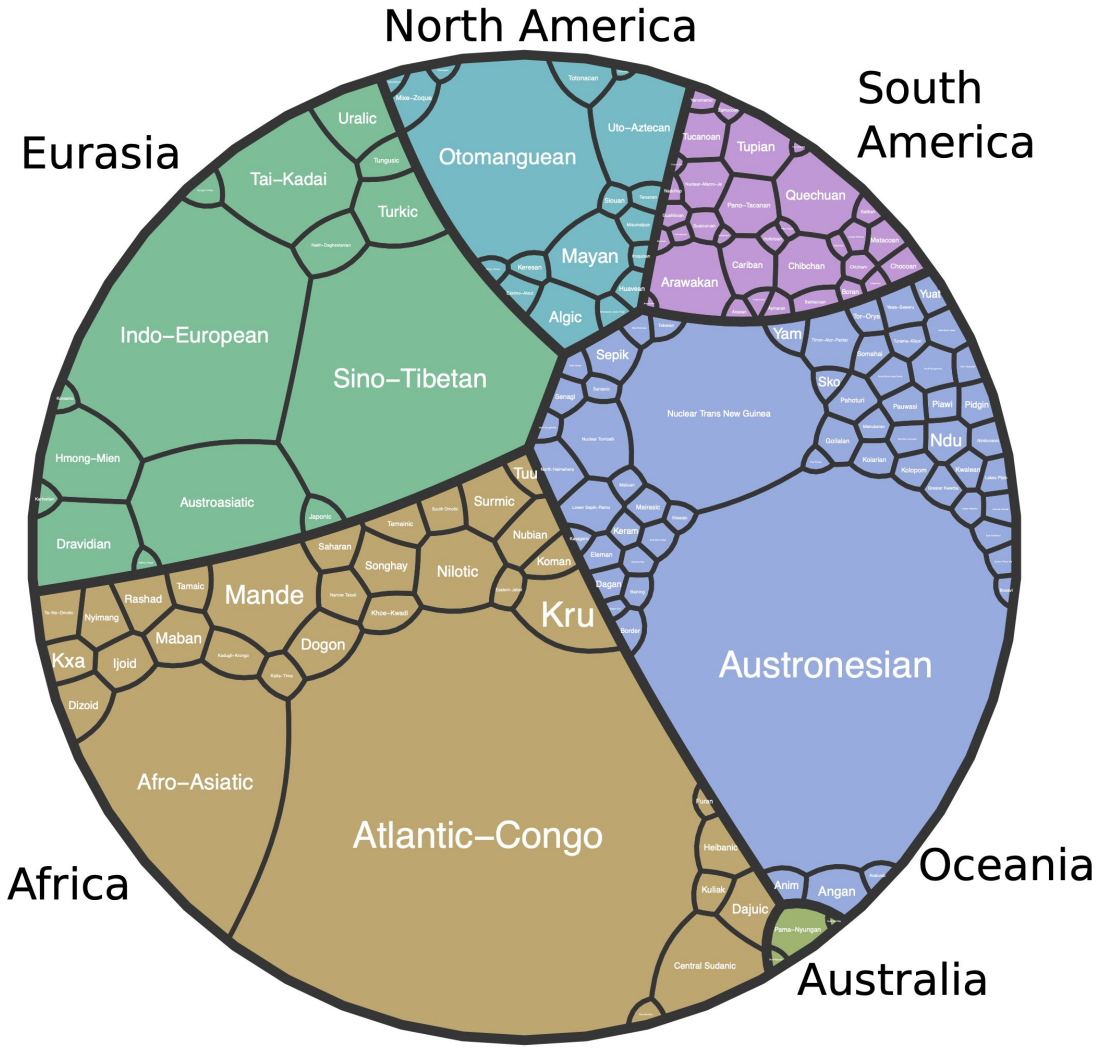
Типометрики

Самира Ферхеес

Семинар по типологии 2021/2022

Слайды: github.com/sverhees/site
> teaching > typology 2021/2022



@SimonJGreenhill, language data: glottolog.org

Типометрики

- Typometrics: From implicational to quantitative universals in word order typology (Gerdes, Kahane & Chen 2021)

Типометрики

- Typometrics: From implicational to quantitative universals in word order typology (Gerdes, Kahane & Chen 2021)
- Проверка универсалии Гринберга про порядок слов с помощью количественных данных (= тексты с синтаксической аннотацией из бд Universal dependencies)

Типологические данные

- Первичные данные хороши тем, что они менее субъективные, чем описания

Типологические данные

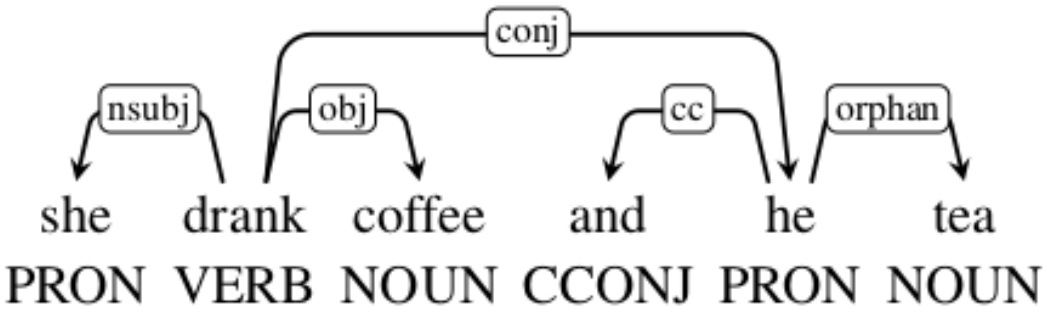
- Первичные данные хороши тем, что они менее субъективные, чем описания
- Но они все равно могут быть необъективными (напр. из-за метода сбора)
- и также не сопоставимыми, как и описания
- и сложно пользоваться не-специалистам

Universal dependencies

- База трибанков – тексты на разных языках, с синтаксической аннотацией по (более менее) единой схеме
- Базовая единица – слово
- Три уровня репрезентации: лемма, тэг для части речи, и features: лексические и грамматические свойства словоформы
- Отношения между единицами в виде Directed Acyclic Graphs (вершина $\rightarrow$ зависимое + тип отношений)

(Nivre et al. 2020)

Universal dependencies



(Nivre et al. 2020: 4038)

Universal dependencies

- С 2014 года
- 157 трибанков, 90 языков
- Пока сильный bias к определенным языкам (в плане присутствия, количества, разнообразия данных)

Universal dependencies

- С 2014 года
- 157 трибанков, 90 языков
- Пока сильный bias к определенным языкам (в плане присутствия, количества, разнообразия данных)
- Проект открытый, любой желающий может контрибьютить свои данные
- Проблемы аннотации активно обсуждаются

(Nivre et al. 2020)

Universal dependencies

Family	Languages
Afro-Asiatic	7
Austro-Asiatic	1
Austronesian	2
Basque	1
Dravidian	2
Indo-European	48
Japanese	1
Korean	1
Mande	1
Mongolic	1
Niger-Congo	2
Pama-Nyungan	1
Sino-Tibetan	3
Tai-Kadai	1
Tupian	1
Turkic	3
Uralic	11
Code-Switching	1
Creole	1
Sign Language	1

(Nivre et al. 2020: 4041)

Universal dependencies

Genre	#	Genre	#
Academic	4	News	98
Bible	10	Non-fiction	57
Blog	17	Poetry	4
Email	2	Reviews	7
Fiction	42	Social	9
Grammar examples	13	Spoken	18
Learner essays	2	Web	9
Legal	22	Wiki	46
Medical	6		

- количество трибанков

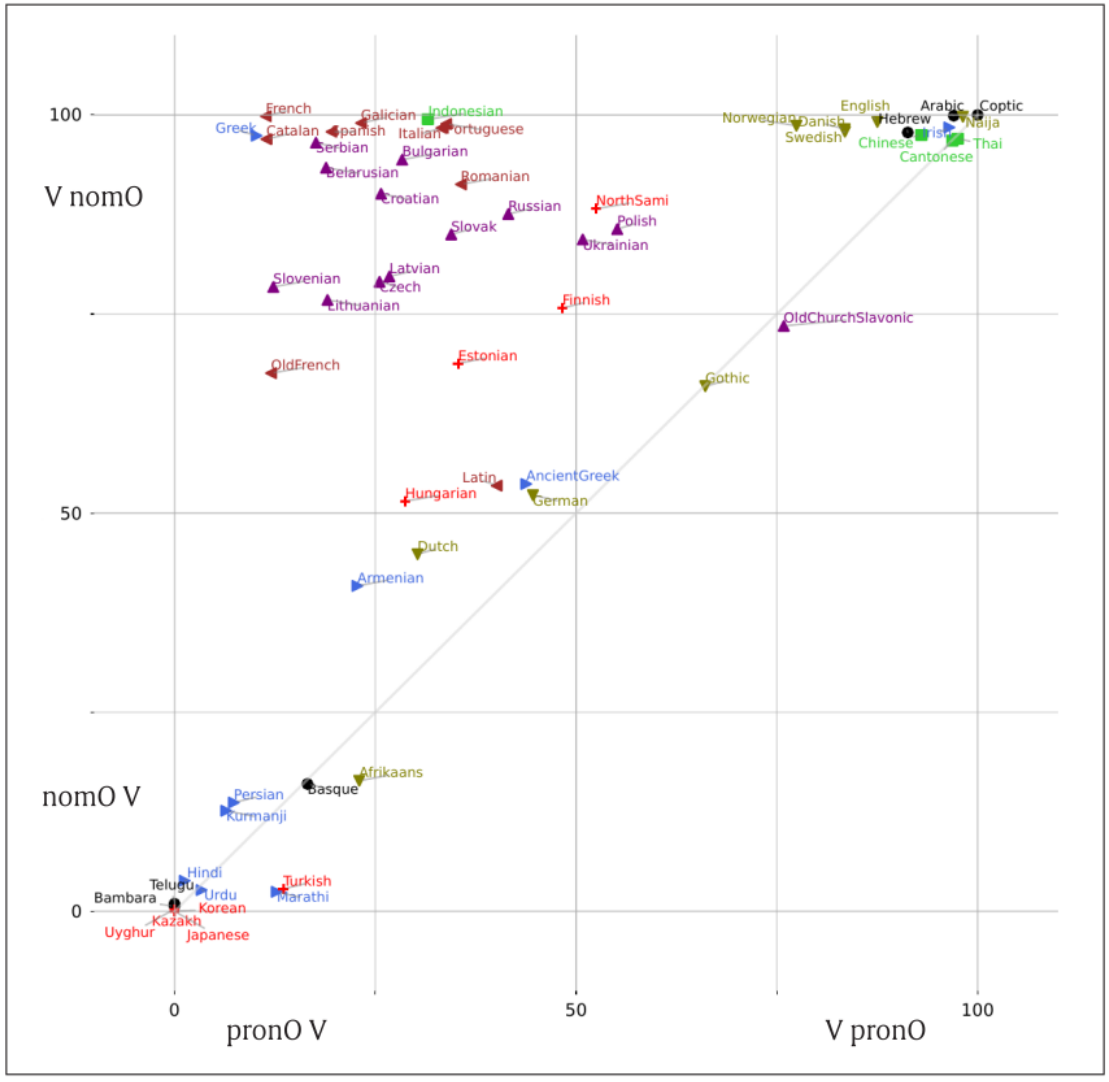
(Nivre et al. 2020: 4042)

Выборка

- Gerdes, Kahane & Chen (2021) взяли всё, что было в UD (хотя цифры отличаются от Nivre et al. (2020))
- Скрыли только слишком маленькие трибанки (< 50 примеров на определенное явление)

Выборка

- Gerdes, Kahane & Chen (2021) взяли всё, что было в UD (хотя цифры отличаются от Nivre et al. (2020))
- Скрыли только слишком маленькие трибанки (< 50 примеров на определенное явление)
- Авторы признают, что их выборка не типологически сбалансированная
- (Не уточняют, сколько языков получилось для каждой представленной семьи и каждого ареала, но названия написаны на графиках)
- Не контролируют жанр (подробнее об этом в Chen & Gerdes (2017))



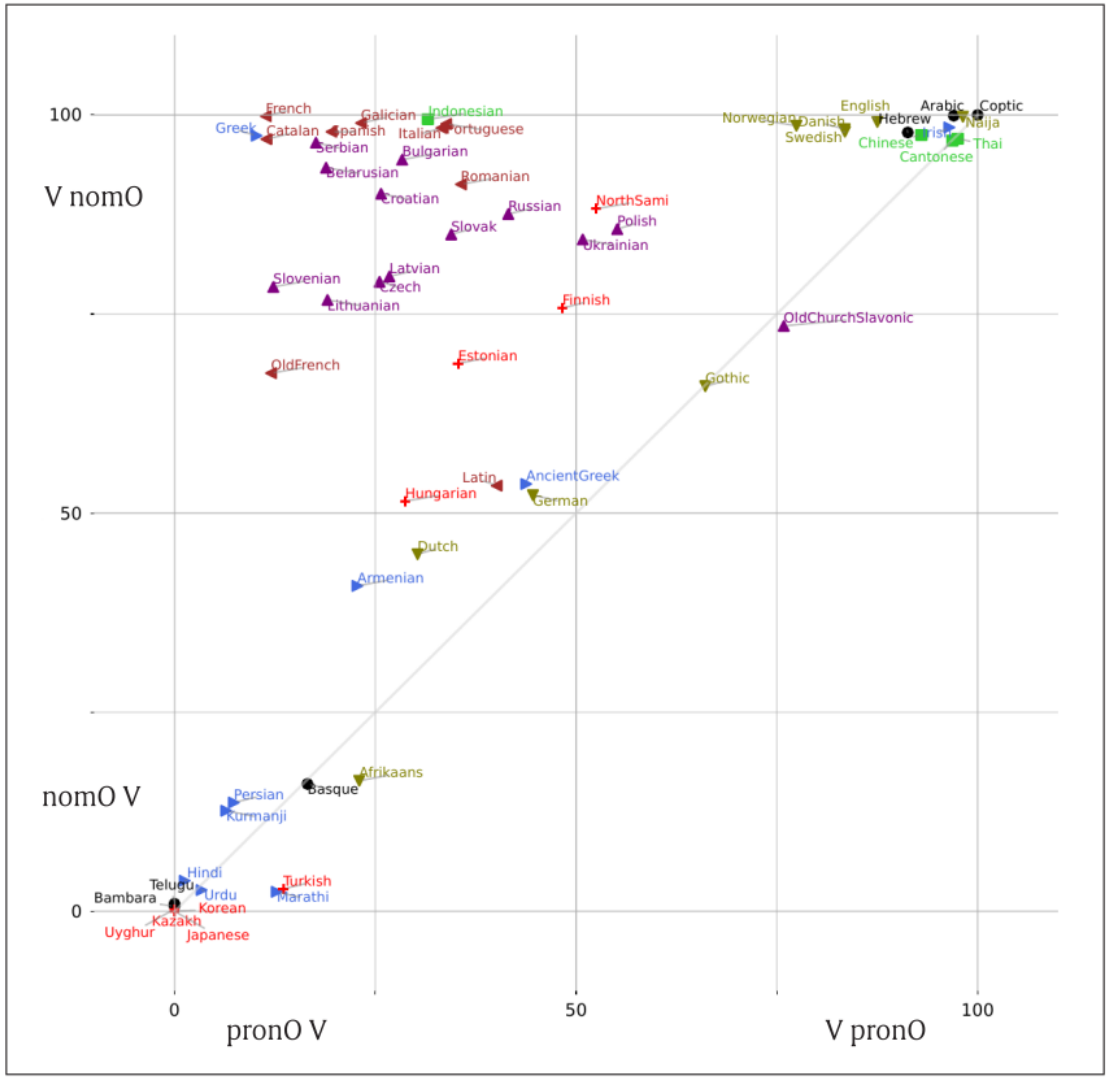
(Gerdes, Kahane & Chen 2021: 3)

Выборка

- Языки раскрашены по разным параметрам:
- Ветки индо-европейской семьи, языковая семья, агглютинативная морфология, и “другие языки”
- Почти выглядит как разнообразная выборка :)

Выборка

- Языки раскрашены по разным параметрам:
- Ветки индо-европейской семьи, языковая семья, агглютинативная морфология, и “другие языки”
- Почти выглядит как разнообразная выборка :)
- Можно ли из этого графика почерпнуть информацию про генеалогическое или ареальное распределение этих признаков и их корреляции?



(Gerdes, Kahane & Chen 2021: 3)

Universal 25

If the pronominal object follows the verb, so does the nominal object. (в архиве, пока без контрпримеров)

(категорическая / квалитативная универсалия)

Universal 25

If the pronominal object follows the verb, so does the nominal object. (в архиве, пока без контрпримеров)

(категорическая / квалитативная универсалия)

Universal 25’

For every language, if the percentage of pronominal objects on the right of the verb is greater than 75%, so is the percentage of nominal objects on the right of the verb. (Gerdes, Kahane & Chen 2021: 6)

(квантитативная универсалия)

Чем квантитативные универсалии отличаются от статистических?

Чем квантитативные универсалии отличаются от статистических?

– Статистические универсалии всё равно содержат квалитативные/категорические утверждения.

напр. *Universal 4. With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.*

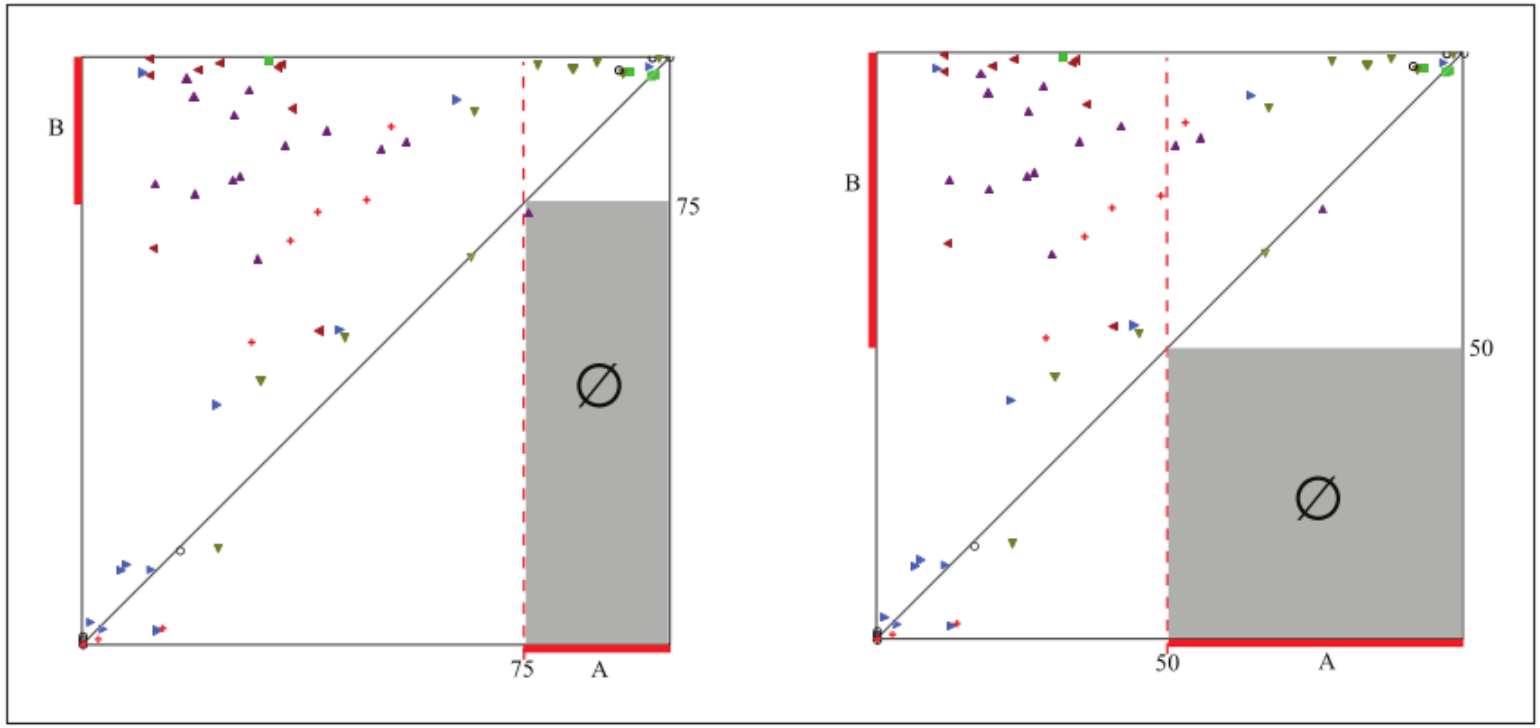
= у языков бывает нормальный / базовый порядок слов, который может быть SOV.

Universal 25’

For every language, if the percentage of pronominal objects on the right of the verb is greater than 75%, so is the percentage of nominal objects on the right of the verb. (Gerdes, Kahane & Chen 2021: 6)

(квантитативная универсалия)

… а почему не 100%? Или 50%?



(Gerdes, Kahane & Chen 2021: 6)

Universal 25’

For every language, if the percentage of pronominal objects on the right of the verb is greater than 75%, so is the percentage of nominal objects on the right of the verb. (Gerdes, Kahane & Chen 2021: 6)

(квантитативная универсалия)

… является ли эта универсалия бидирективной?

Universal 25’

For every language, if the percentage of pronominal objects on the right of the verb is greater than 75%, so is the percentage of nominal objects on the right of the verb. (Gerdes, Kahane & Chen 2021: 6)

(квантитативная универсалия)

… является ли это абсолютной или статистической универсалией?

Inequality Universal

Almost every language has a higher proportion of nominal objects than of pronominal objects on the right of the verb. (Gerdes, Kahane & Chen 2021: 7)

Inequality Universal

Almost every language has a higher proportion of nominal objects than of pronominal objects on the right of the verb. (Gerdes, Kahane & Chen 2021: 7)

(исключения: африкаанс, турецкий, маратхи, старославянский)

Inequality Universal

Almost every language has a higher proportion of nominal objects than of pronominal objects on the right of the verb. (Gerdes, Kahane & Chen 2021: 7)

(исключения: африкаанс, турецкий, маратхи, старославянский)

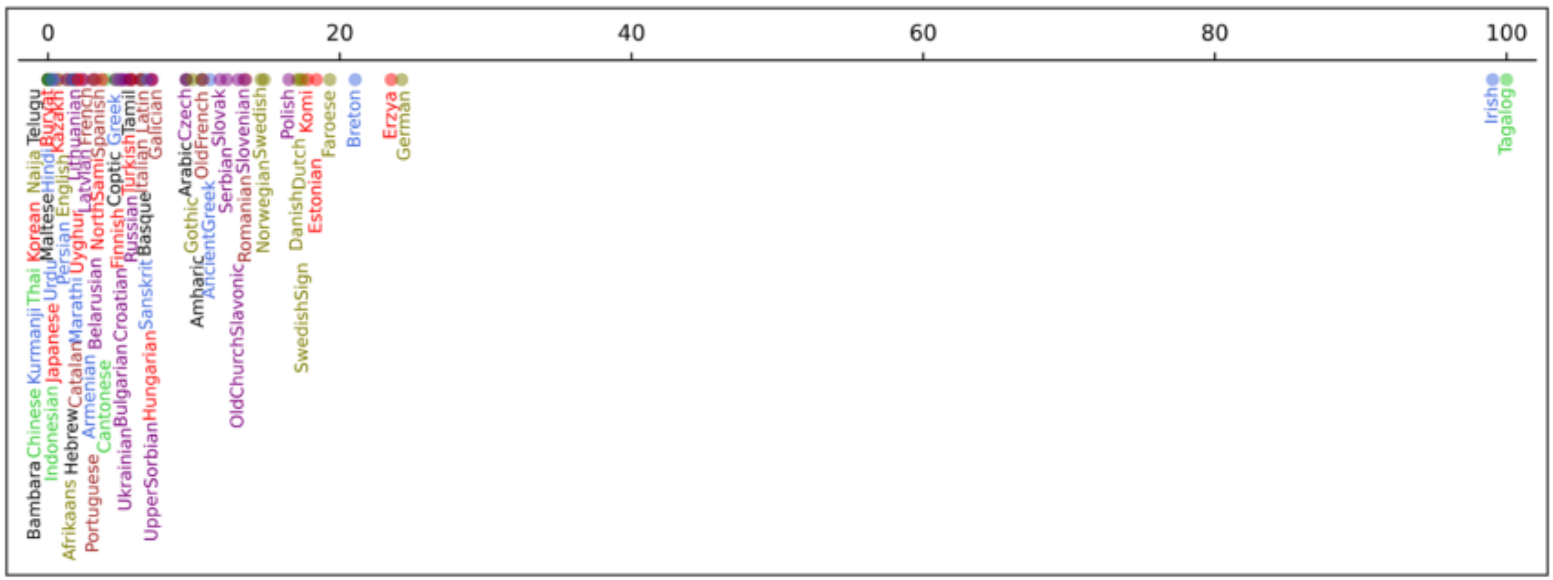
Почему местоименные объекты ведут себя по-другому чем именные?

Порядки других составляющих

- Показывают тенденции языков к head-initial или head-final порядок
- И вариативность между разными конструкциями

Порядки других составляющих

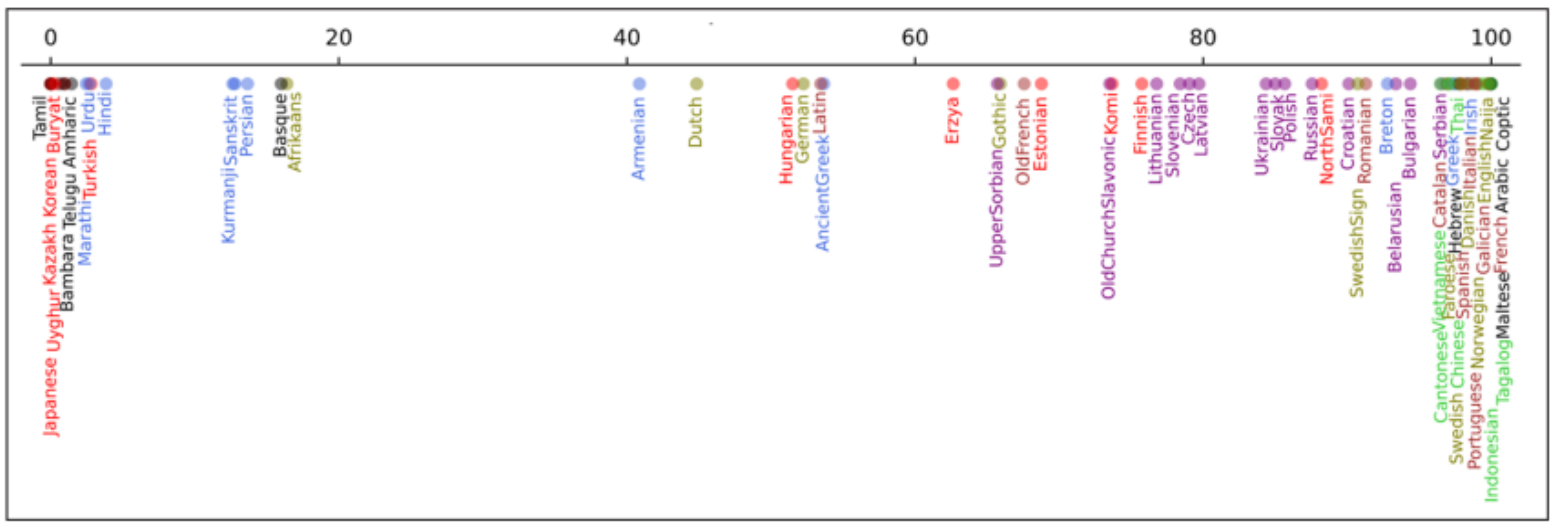
Количество случаев, когда местоименный *субъект* находился после глагола



(Gerdes, Kahane & Chen 2021: 17)

Свободные и смешанные порядки

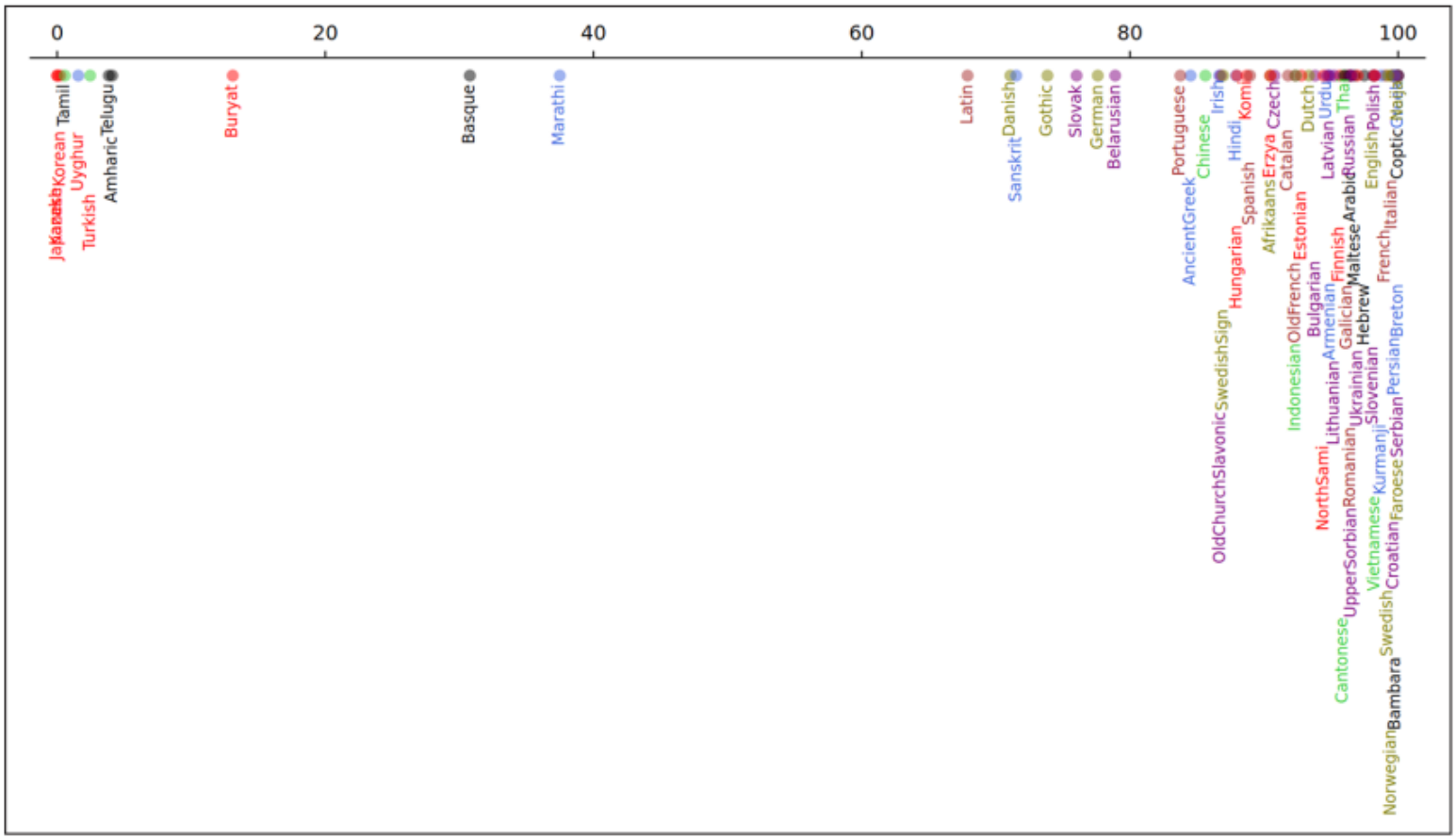
Количество случаев, когда номинальный объект находился после глагола



(Gerdes, Kahane & Chen 2021: 15)

Порядки других составляющих

Количество случаев, когда глагольный комплемент находился после вершины



(Gerdes, Kahane & Chen 2021: 18)

Порядки других составляющих

"More generally it appears that light dependents tend to be more on the left than heavier dependents if we consider pronouns to be lighter than noun phrases, noun phrases to be lighter than clauses, and adverbial modifiers to be lighter than adverbial clauses." (Gerdes, Kahane & Chen 2021: 18)

Порядки других составляющих

“More generally it appears that light dependents tend to be more on the left than heavier dependents if we consider pronouns to be lighter than noun phrases, noun phrases to be lighter than clauses, and adverbial modifiers to be lighter than adverbial clauses.” (Gerdes, Kahane & Chen 2021: 18)

(вспомним Hawkins (1990))
(но имеем в виду bias выборки)

Проблема вершины

- Не всегда понятно, что является вершиной
- Спорные составляющие (адложные фразы, вспомогательные глаголы, копулы)
- Спорные языки (напр. предлоги/ковербы в китайских языках (Gerdes, Kahane & Chen 2021: 14–15))

SUD

- Gerdes, Kahane & Chen (2021) используют свою трансформацию UD аннотации: Surface syntactic UD
- UD и SUD аннотации в принципе совместимы, и одну легко конвертировать в другую (и обратно)
- Но схемы по другому идентифицируют вершины (UD: content words, a SUD: funtional heads)
- Напр. *Peter talked to*$_{SUD}$ *Mary*$_{UD}$
- В каких-то случаях это влияет на позицию составляющих

Типометрики

- Типометрика – типология на основе (относительно) больших данных, вместо квалитативных, напр. из дескриптивной литературы
- В отличие от квантитативной типологии, в типометрике не стремятся к типологически сбалансированным выборкам

Типометрики

- Типометрика – типология на основе (относительно) больших данных, вместо квалитативных, напр. из дескриптивной литературы
- В отличие от квантитативной типологии, в типометрике не стремятся к типологически сбалансированным выборкам
- В случае исследования Gerdes, Kahane & Chen (2021), преобладают индо-европейские языки
- Поэтому выводы нужно считать предварительными

Типометрики

- Зато легко фальсифицировать исследование когда появится больше языков в UD
- И таким образом можно сформулировать всё более точные универсалии
- Решает некоторые проблемы применимости универсалий к конкретным языкам (что такое “базовый порядок слов”)
- Но не решает проблемы определения вершины
- (Возможно даже делает хуже в связи с разными авторскими подходами к аннотации в UD)

Типометрики

- В статье пока не предлагают интерпретацию результатов для языков из выборки
- И особо не сравнивают свои результаты с результатами квантитативной типологии (кроме обсуждения head-final vs. head-initial (Gerdes, Kahane & Chen 2021: 14))
- Хотя как раз было бы интересно посмотреть, какие квантитативные метрики стоят за квалитативными утверждениями в других типологических работах
- Ср. например главу в ВАЛСе Order of object and verb (Dryer 2013)

Литература I


Chen, Xinying & Kim Gerdes. 2017. Classifying languages by dependency structure. Typologies of delexicalized universal dependency treebanks. In *Proceedings of the fourth international conference on dependency linguistics (depling 2017)*, 54–63.


Dryer, Matthew S. 2013. Order of object and verb. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/83.

Литература II

- Gerdes, Kim, Sylvain Kahane & Xinying Chen. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa* 6(1).
- Hawkins, John A. 1990. A parsing theory of word order universals. *Linguistic inquiry* 21(2). 223–261.
- Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers & Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.