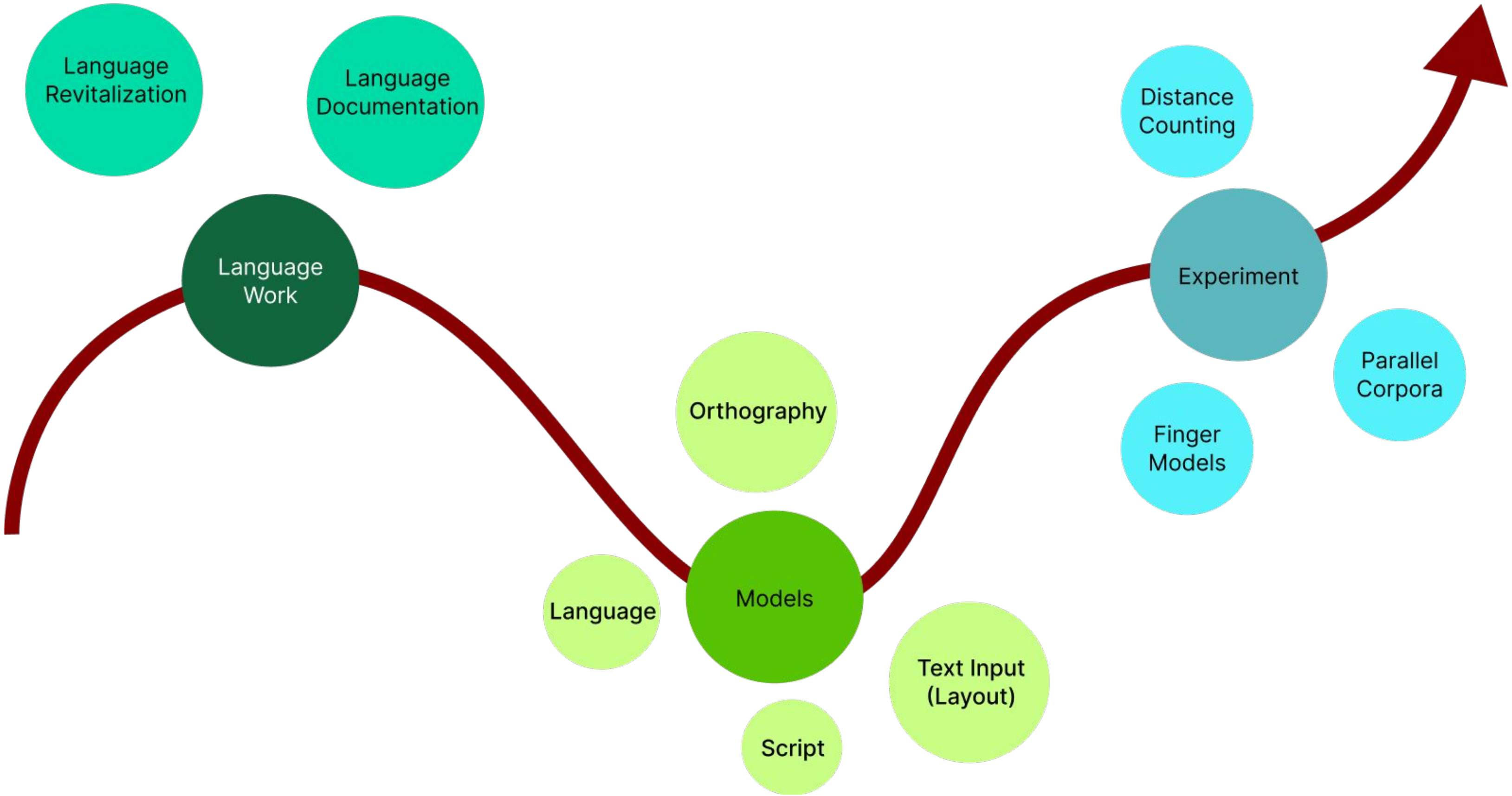
Language, Script, Orthography, and Text-input

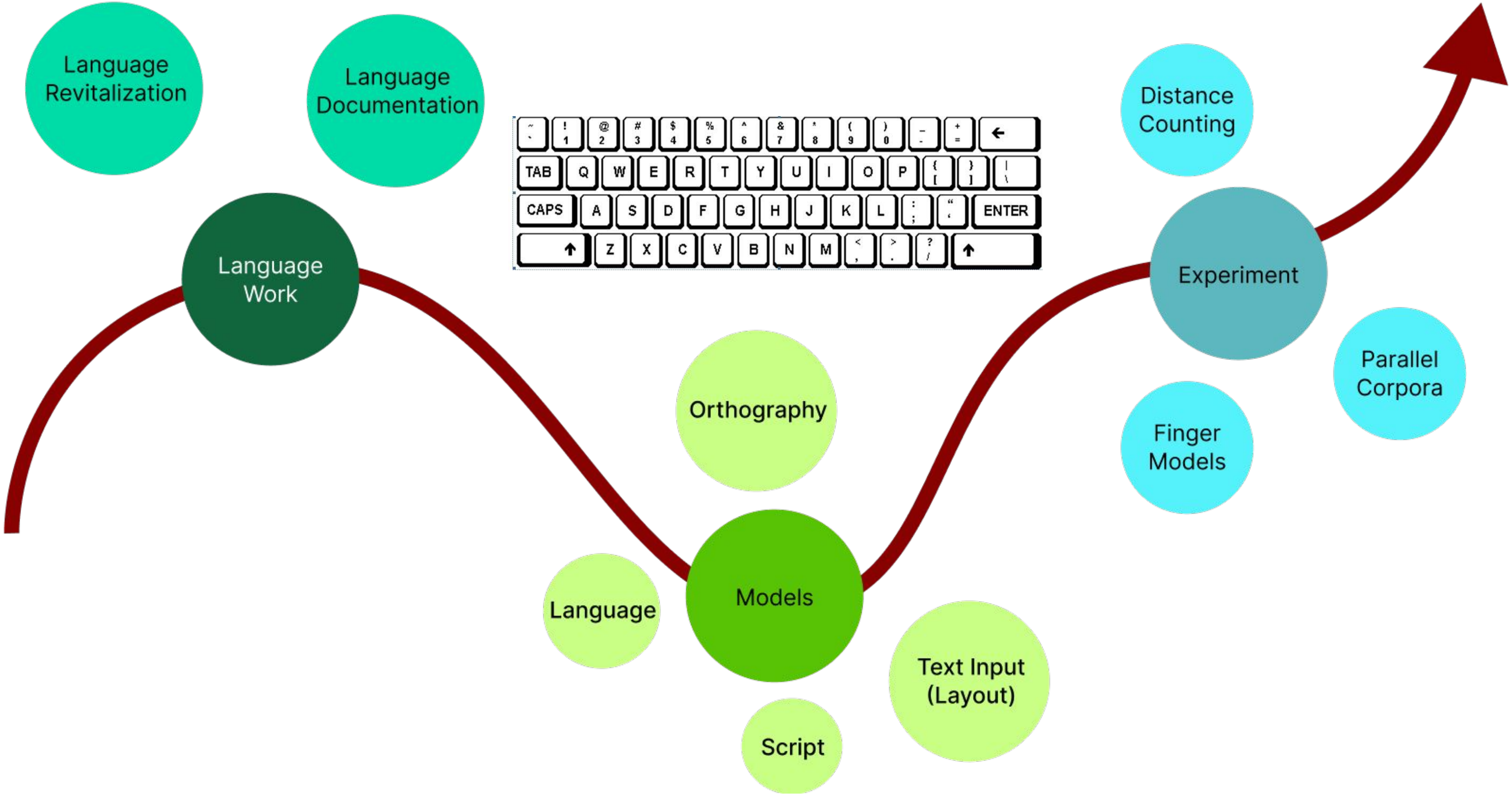
Rating Text-input Difficulty Across Languages

Hugh Paterson III
15 November 2022

TopicMap


Language Revitalization
Language Documentation
Language Work
Orthography
Language
Models
Script
Text Input (Layout)
Distance Counting
Experiment
Parallel Corpora
Finger Models

TopicMap



Language Revitalization

Language Documentation

Language Work

Orthography

Language

Models

Script

Text Input (Layout)

Distance Counting

Experiment

Parallel Corpora

Finger Models

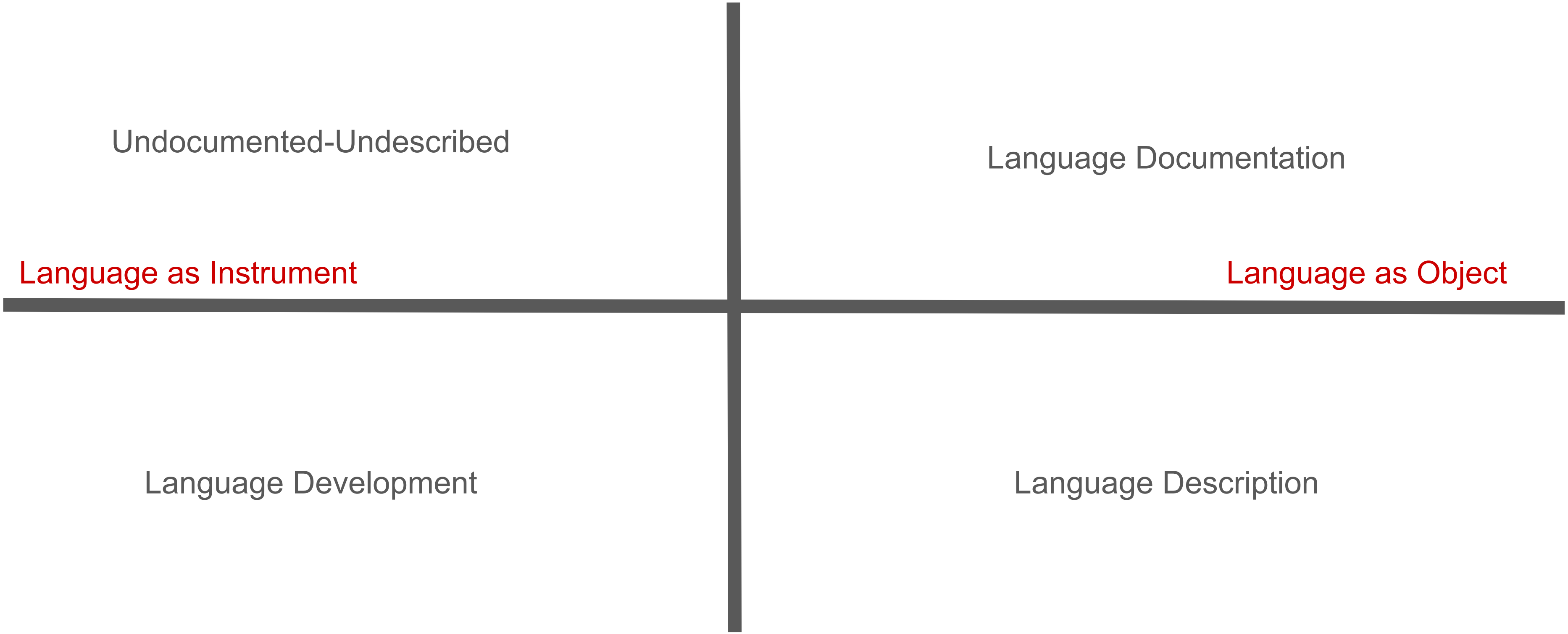
The Economics of Linguistic Exchanges

Language as Instrument

Language as Object

Bourdieu, Pierre. 1977. “The economics of linguistic exchanges.” *Social Science Information* 16(6). 645–668. doi:10.1177/053901847701600601

Broad Categories for Language Resources



Undocumented-Undescribed

Language Documentation

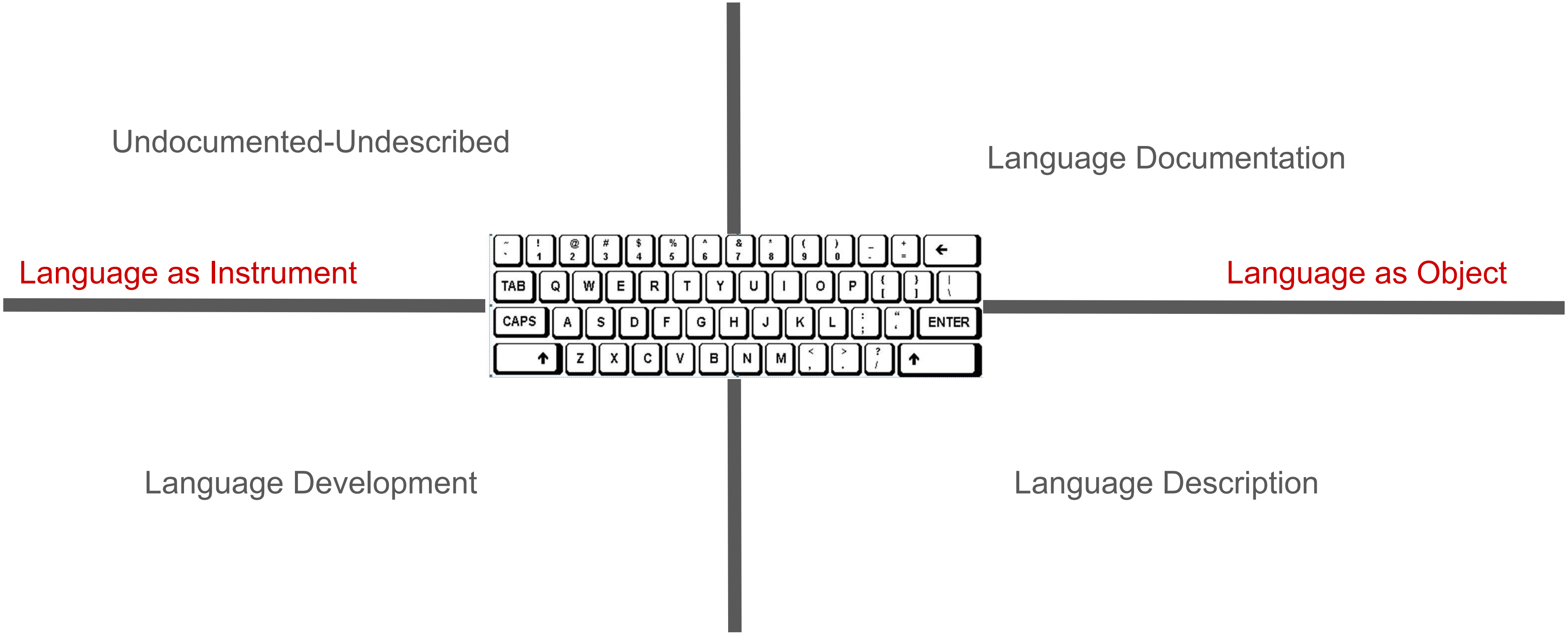
Language as Instrument

Language as Object

Language Development

Language Description

Broad Categories for Language Resources



~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 _ - + = ←
TAB Q W E R T Y U I O P { [}] | \
CAPS A S D F G H J K L : ; " ' ENTER
↑ Z X C V B N M < , > . ? / ↑

Undocumented-Undescribed

Language Documentation

Language as Instrument

Language as Object

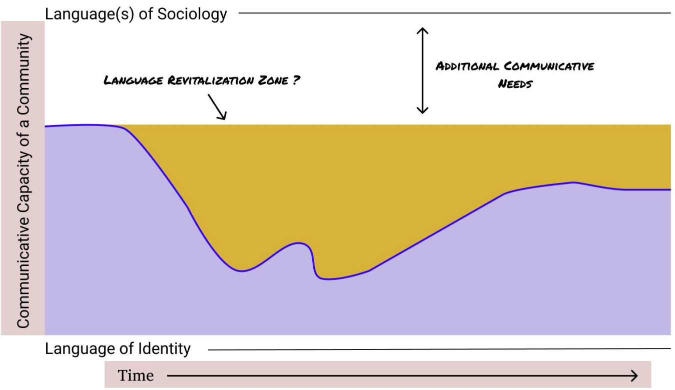
Language Development

Language Description

Broad Categories for Language Resources

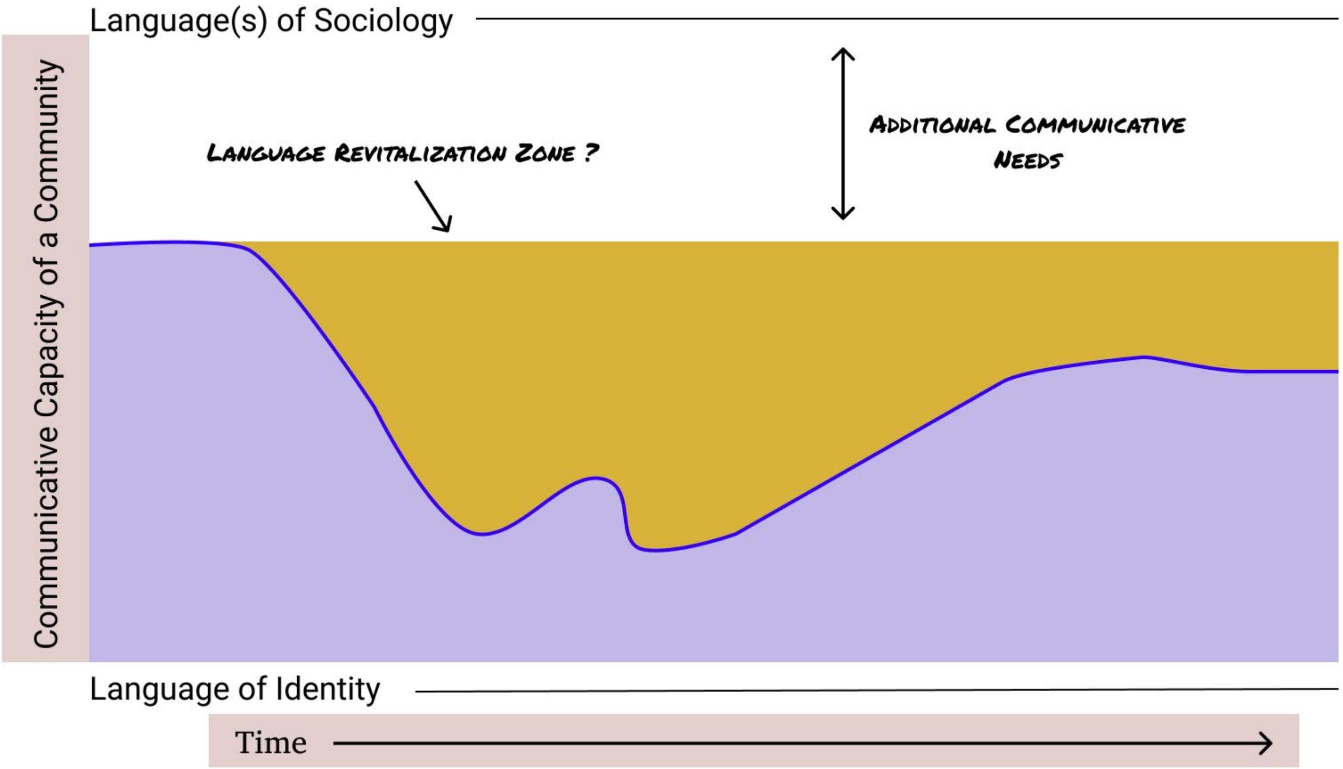
Undocumented-Undescribed

Language Documentation



Language Description

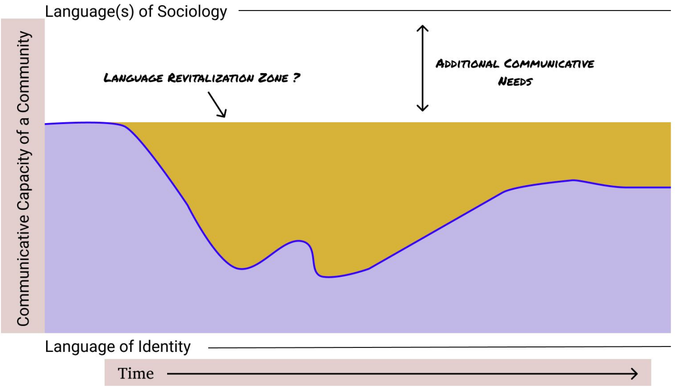
Broad Categories for Language Resources


Language(s) of Sociology
Additional Communicative Needs
Language Revitalization Zone ?
Communicative Capacity of a Community
Language of Identity
Time

Broad Categories for Language Resources

Undocumented-Undescribed

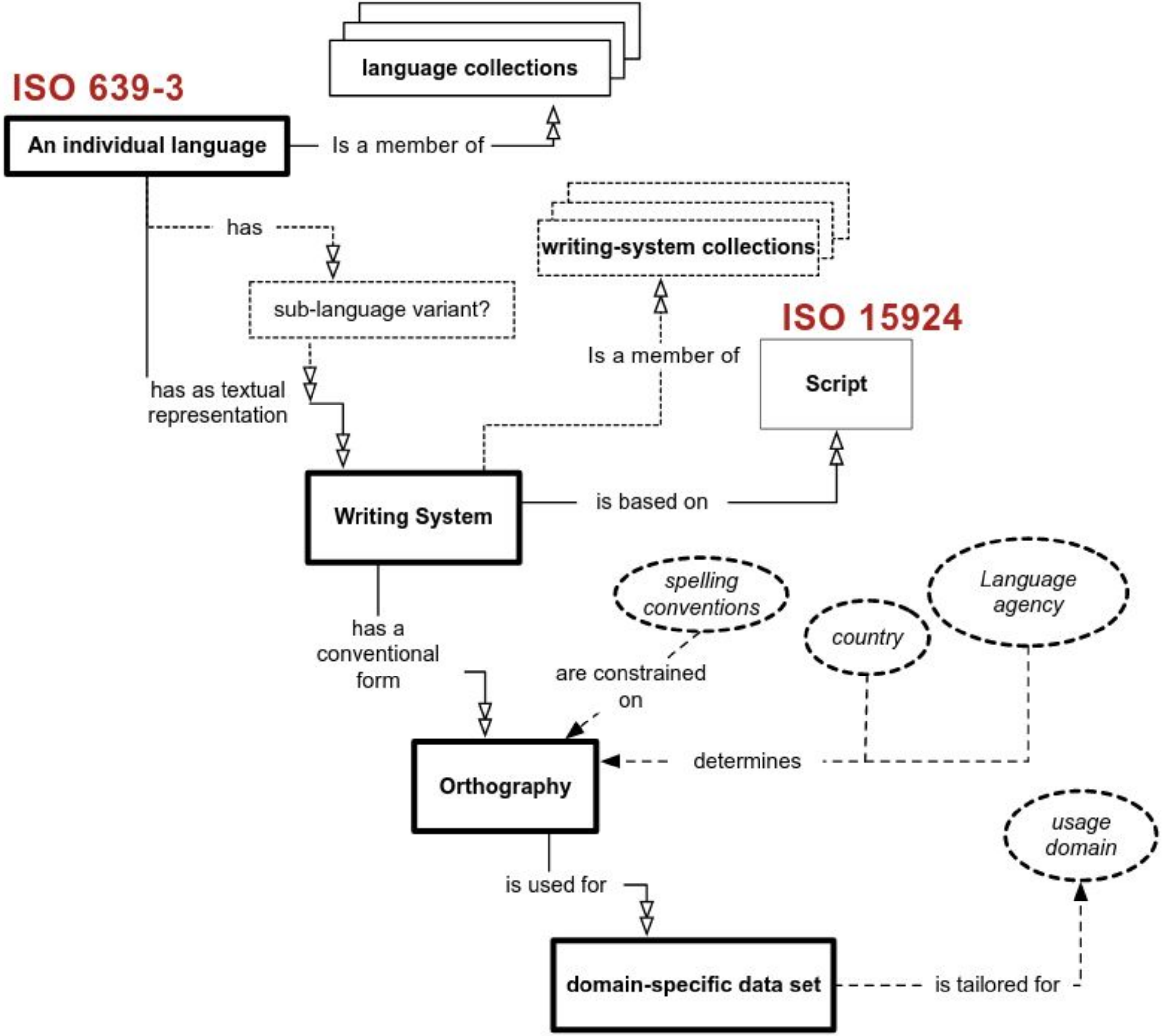
Language Documentation



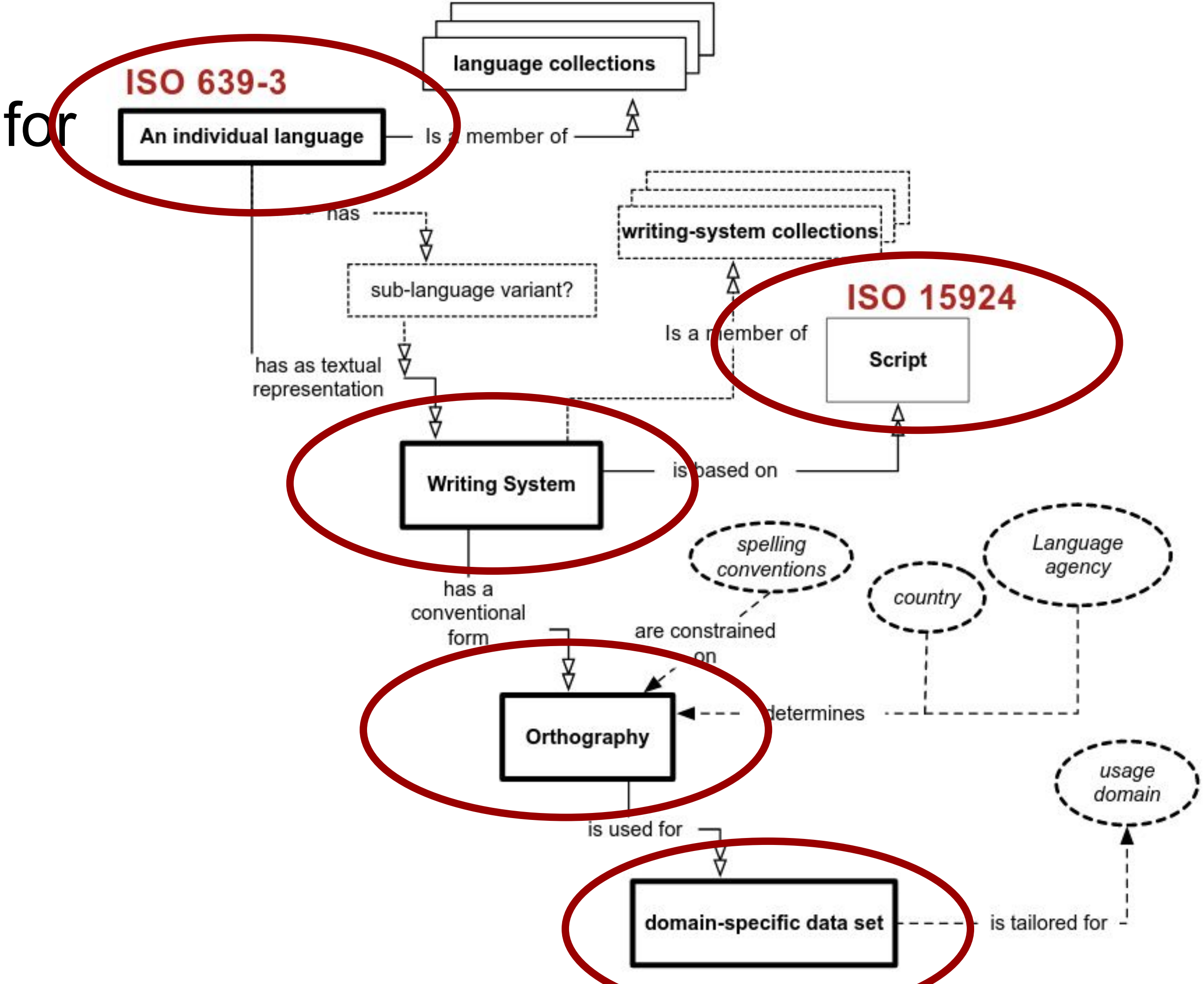
Language Description

A Useful Model for Languages and Scripts

A Useful Model for Languages and Scripts



A Useful Model for Languages and Scripts

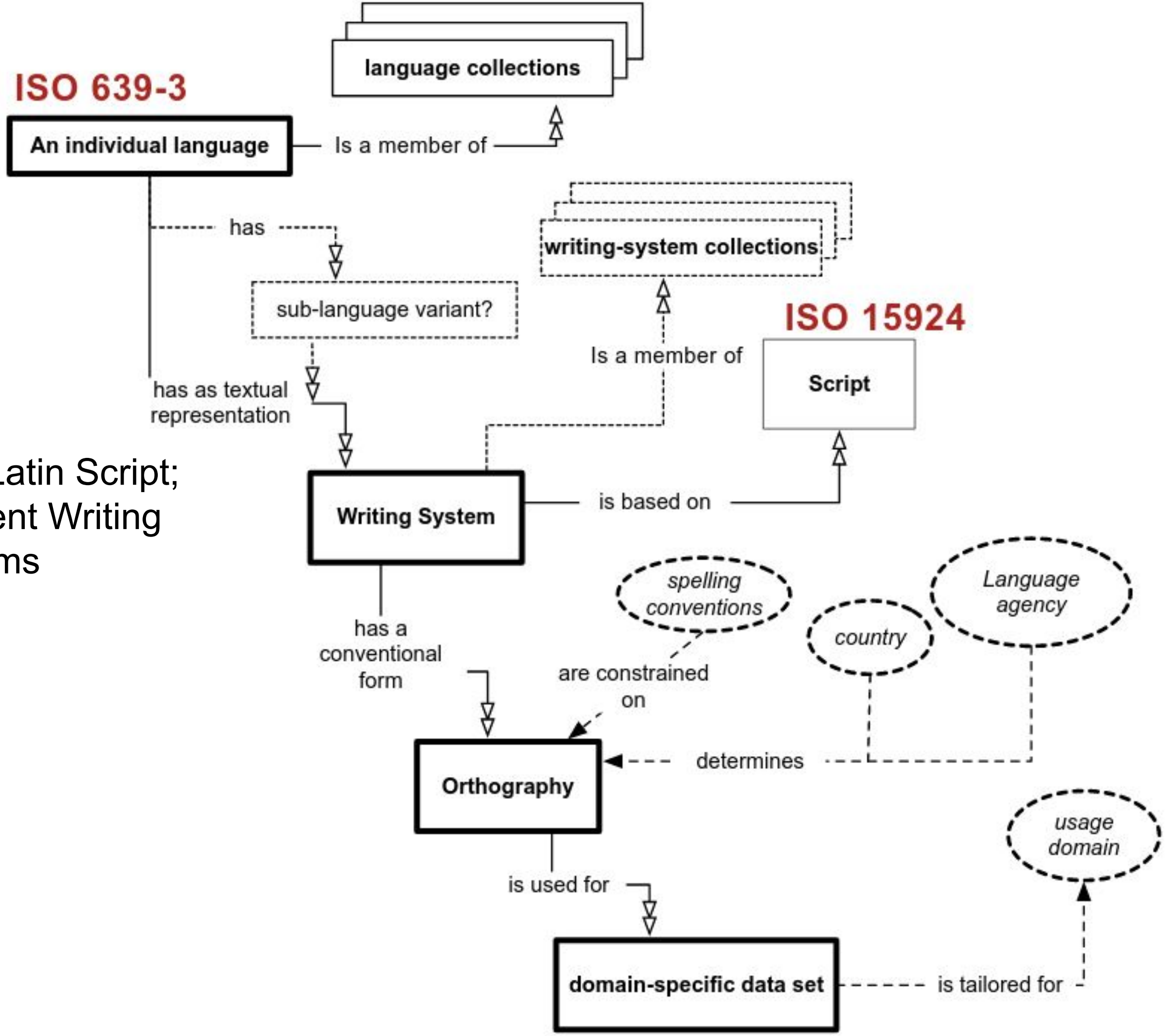


A Useful Model for Languages and Scripts

English: ____________

German: äÄ, öÖ, üÜ, ß

} Both Latin Script; Different Writing Systems



A Useful Model for Languages and Scripts

English: ___________

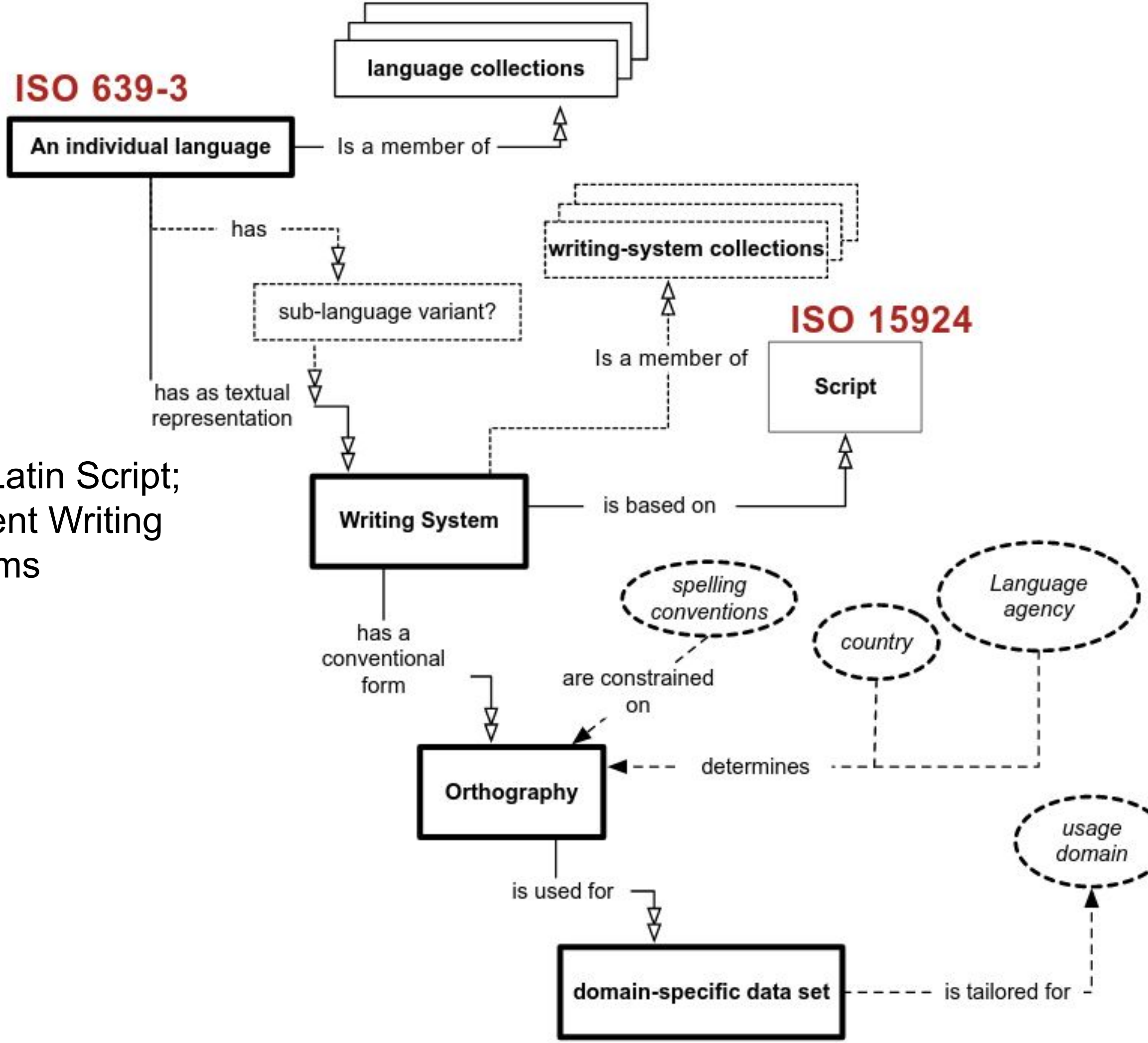
German: äÄ, öÖ, üÜ, ß

} Both Latin Script; Different Writing Systems

British English
American English

German prior to 1996
German post 1996
German in Switzerland

}



A Useful Model for Languages and Scripts

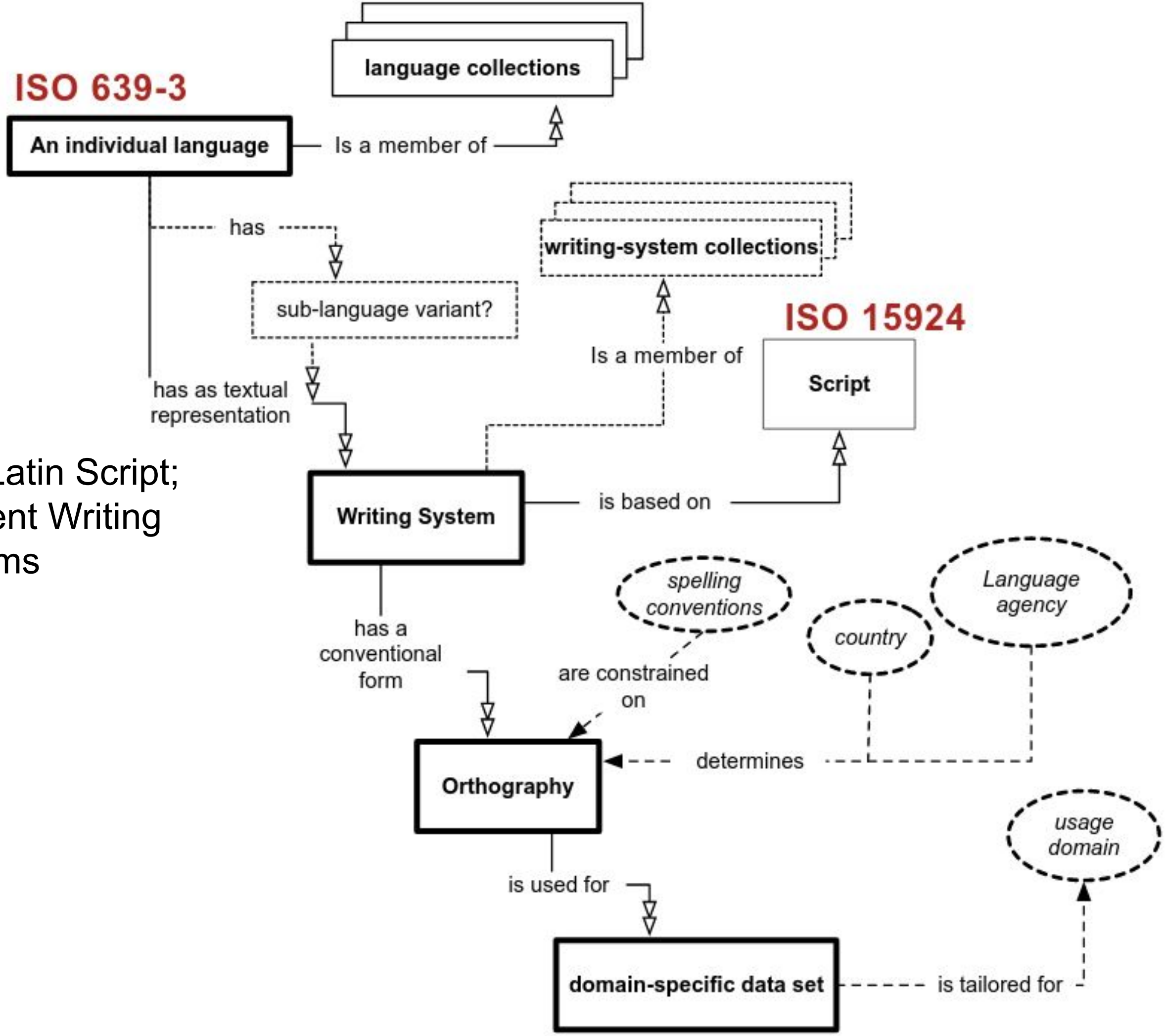
English: ___________

German: äÄ, öÖ, üÜ, ß

} Both Latin Script; Different Writing Systems

British English
American English

German prior to 1996
German post 1996
German in Switzerland }



A Useful Model for Languages and Scripts

English: ____________

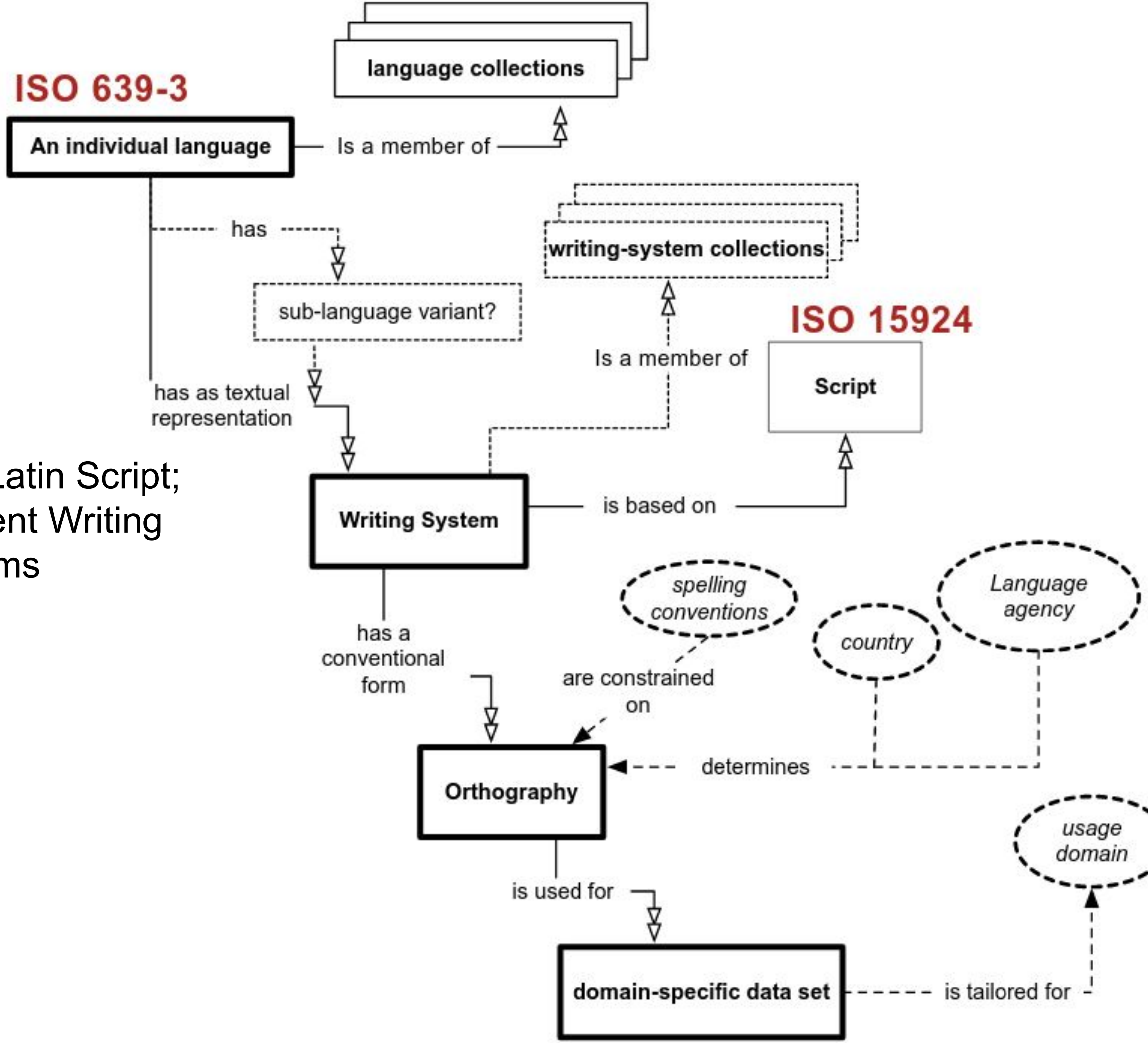
German: äÄ, öÖ, üÜ, ß

} Both Latin Script; Different Writing Systems

British English
American English

German prior to 1996
German post 1996
German in Switzerland

Newspaper Articles
Library Records
Interlinear-Glossed Texts



Text Input

Text Input

Speech-to-text

Predictive models—T9

Swipe based

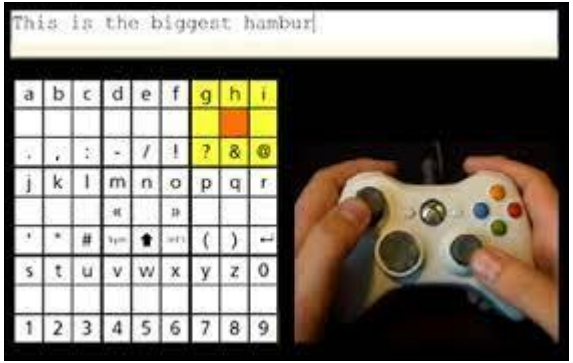
Eye gaze

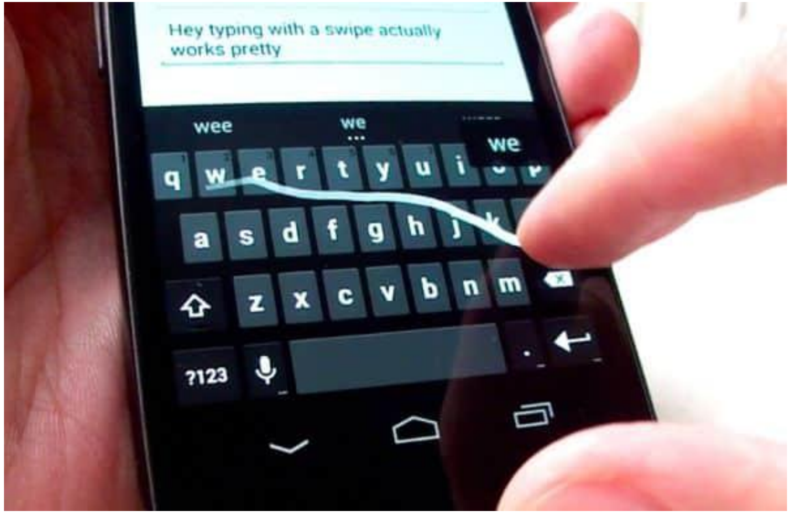
Thumb based cell-phones

Stylus based

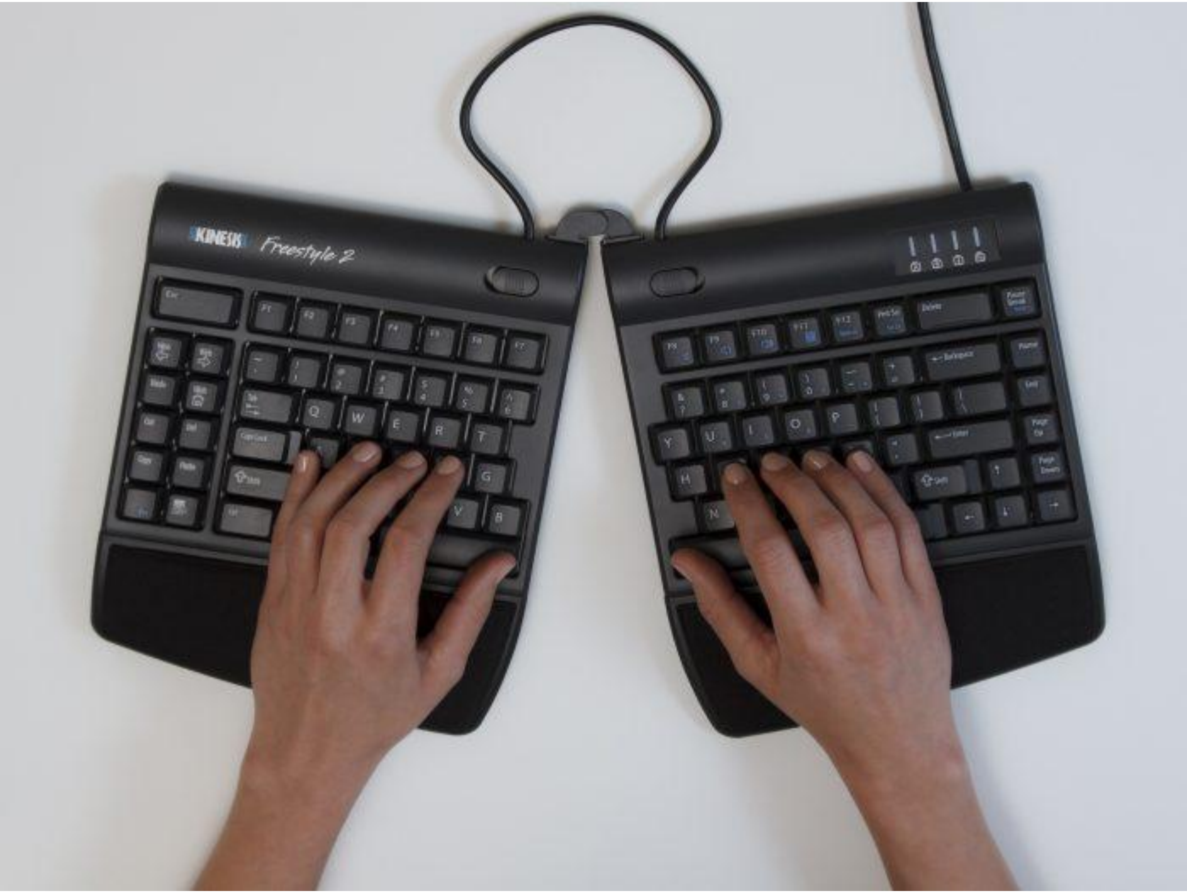
Controller based

OMG LOL KWL. G2G SOZ. TTYL. LU T8


This is the biggest hambur

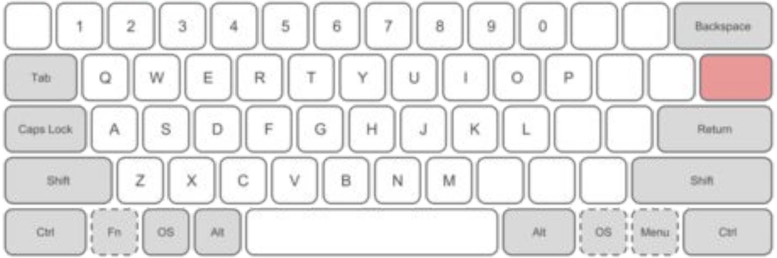

Hey typing with a swipe actually works pretty

Keyboards

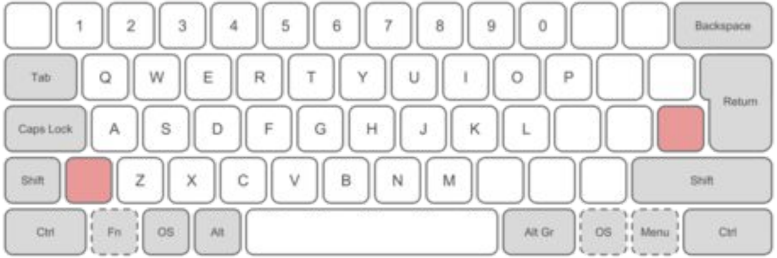


Keyboards

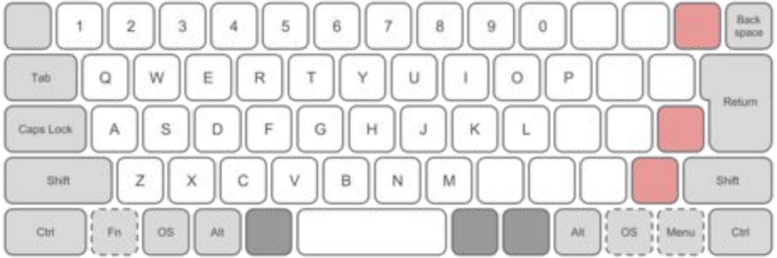
ANSI 米国国家規格協会


1 2 3 4 5 6 7 8 9 0 Backspace
Tab Q W E R T Y U I O P
Caps Lock A S D F G H J K L Return
Shift Z X C V B N M Shift
Ctrl Fn OS Alt Alt OS Menu Ctrl

ISO 国際標準化機構


1 2 3 4 5 6 7 8 9 0 Backspace
Tab Q W E R T Y U I O P Return
Caps Lock A S D F G H J K L
Shift Z X C V B N M Shift
Ctrl Fn OS Alt Alt Gr OS Menu Ctrl

JIS 日本工業規格


1 2 3 4 5 6 7 8 9 0 Back space
Tab Q W E R T Y U I O P Return
Caps Lock A S D F G H J K L
Shift Z X C V B N M Shift
Ctrl Fn OS Alt Alt OS Menu Ctrl

A Reference to the Abstract

Would a certain text be easier to type in:

- English [eng],
- French [fra], or
- Eastern Dan [dnj]?

Dear brothers and sisters, when troubles come your way, consider it an opportunity for great joy.

Estimez-le comme une parfaite joie, mes frères, quand vous serez en butte à diverses tentations, sachant que l’épreuve de votre foi produit la patience.

-A pö 'a- wo "dhʋ̈ bha- 'klɔɔ- -mü =dhɛ -a -dhɛa -bha =dhɛ -yö kë ka "yaan =dhɛ 'wɔn "gbɩɩgbɩɩ- -nu bha =në- =gban =ya 'kun bho ka 'gü ka -bha 'dhang -bho Yesu "dhiʋ̈ -sü bha- 'gü, 'yö dho ka gba zuësɛadhɛ 'ka.

Tone Orthographies in Latin Scripts I

Roberts (2011) *A Tone Orthography Typology*.

— a must read for anyone working on developing a tonal orthography in a latin script.

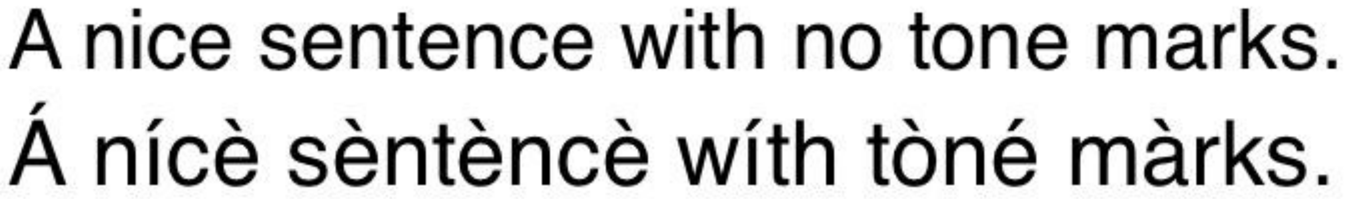
- **Domain**
- **Target**
- **Symbol**
- **Position**
- **Density**
- **Depth**

Density: Some orthographies represent tone exhaustively, that is to say every tone bearing unit carries a symbol for tone, so tone diacritic density is 100%. (p.92)

Tone diacritic density is precisely quantifiable by calculating the number of tone diacritics in a natural text (100 word sample) as a percentage of the number of tone bearing units (Bird 1999: 89). (Roberts p. 90)

-A pö 'a- wo "dhʋ̈ bha- 'klɔɔ- -mü =dhɛ -a -dhɛa -bha =dhɛ -yö kë ka "yaan =dhɛ 'wɔn "gbɯgbɯ- -nu bha =në- =gban =ya 'kun bho ka 'gü ka -bha 'dhang -bho Yesu "dhiʋ̈ -sü bha- 'gü, 'yö dho ka gba zuësɛadhɛ 'ka.

Tone Orthographies in Latin Scripts II

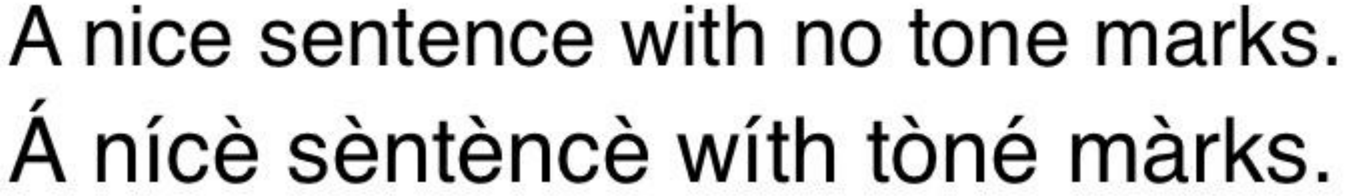


Zero Tone Marking A nice sentence with no tone marks.

Exhaustive Tone Marking Á nícè sèntèncè wíth tòné màrks.

Tone Orthographies in Latin Scripts II

What about Tone Melodies/Tonal Patterns?



Zero Tone Marking A nice sentence with no tone marks.

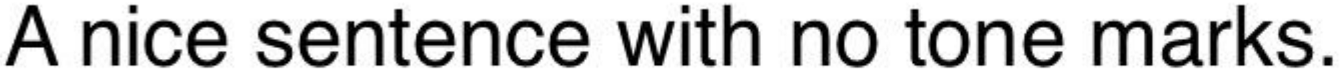
Exhaustive Tone Marking Á nícè sèntèncè wíth tòné màrks.

Tone Orthographies in Latin Scripts II

What about Tone Melodies/Tonal Patterns?

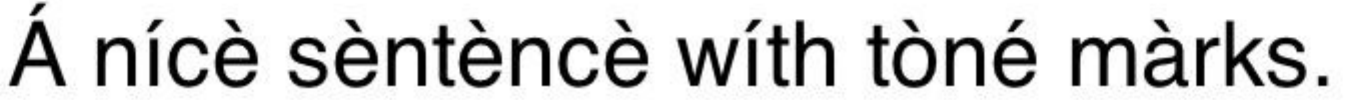
This analytical approach argues that the pitch “phrase” at the domain of the word operates as a single phonological unit. This method then comes in conflict with methods of analysis operating as if the domain of interpretation of pitch is solely the vowel.

Zero Tone Marking



A nice sentence with no tone marks.

Exhaustive Tone Marking

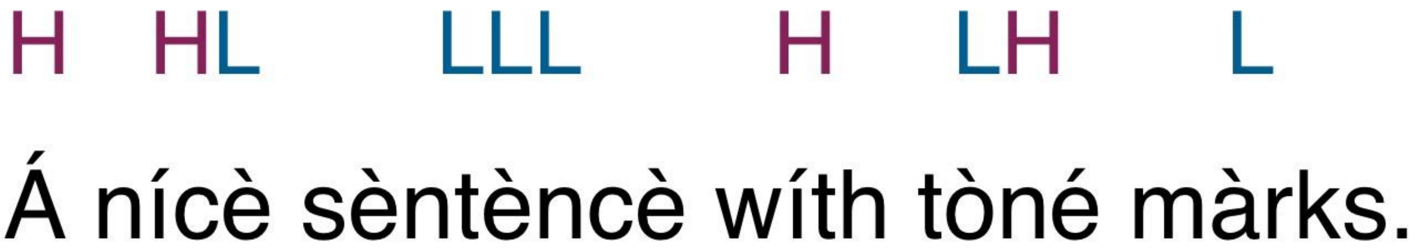


Á nícè sèntèncè wíth tòné màrks.

Tone Orthographies in Latin Scripts II

What about Tone Melodies/Tonal Patterns?

This analytical approach argues that the pitch “phrase” at the domain of the word operates as a single phonological unit. This method then comes in conflict with methods of analysis operating as if the domain of interpretation of pitch is solely the vowel.



H HL LLL H LH L

Á nícè sèntèncè wíth tòné màrks.

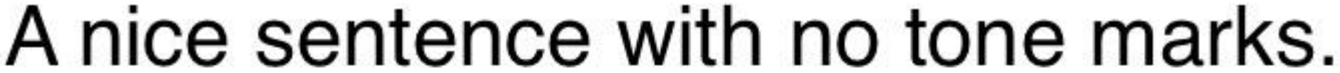
Tone Orthographies in Latin Scripts II

What about Tone Melodies/Tonal Patterns?

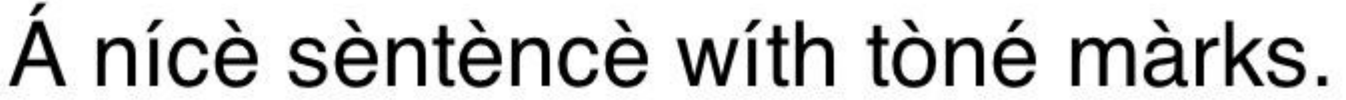
This analytical approach argues that the pitch “phrase” at the domain of the word operates as a single phonological unit. This method then comes in conflict with methods of analysis operating as if the domain of interpretation of pitch is solely the vowel.

Bird’s method is not a great method because it is phonologically accurate, but rather because it is consistently countable.

It represents a structuralist approach to phonemic analysis applied to orthography.

Zero Tone Marking 

A nice sentence with no tone marks.

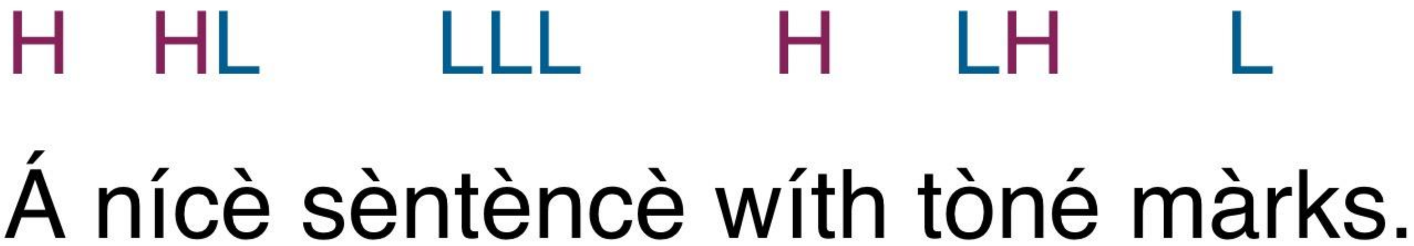
Exhaustive Tone Marking 

Á nícè sèntèncè wíth tòné màrks.

Tone Orthographies in Latin Scripts III

What happens if one is analyzing a language where a pattern occurs describable like:

“No two high-pitches can occur sequentially”....

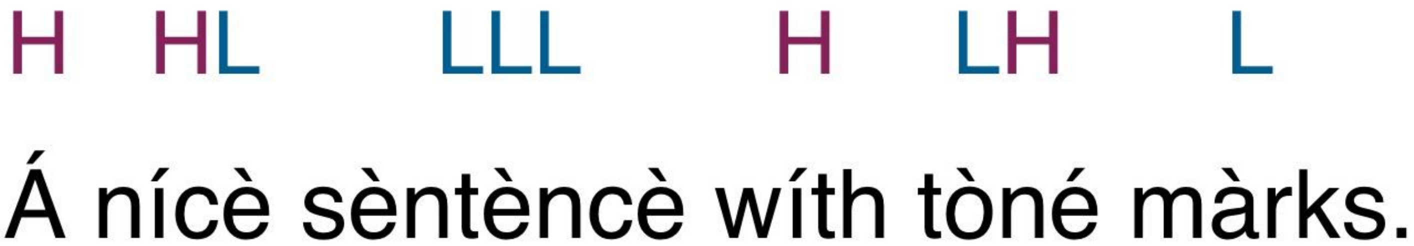


Á nícè sèntèncè wíth tòné màrks.

Tone Orthographies in Latin Scripts III

What happens if one is analyzing a language where a pattern occurs describable like:
“No two high-pitches can occur sequentially”....

Depth: (Roberts)

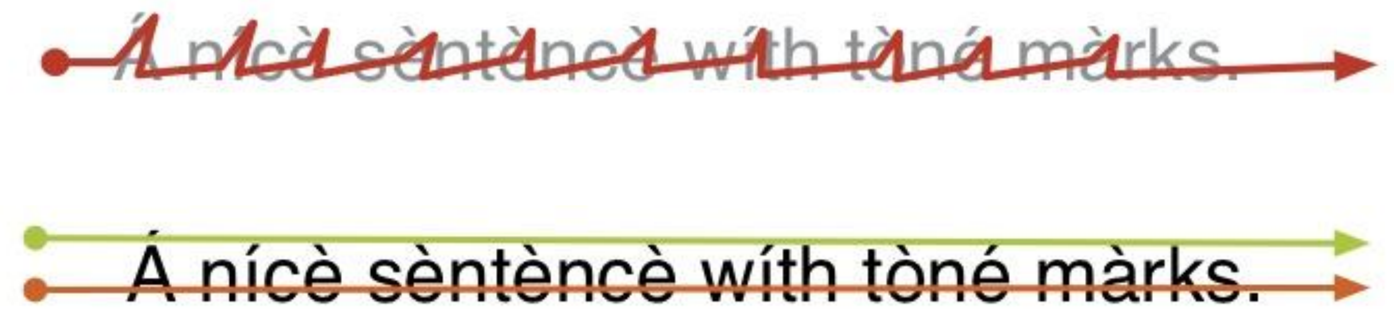


H HL LLL H LH L

Á nícè sèntèncè wíth tòné màrks.

Hard to Use… (the cognitive domain)

What does “hard to use” really mean?

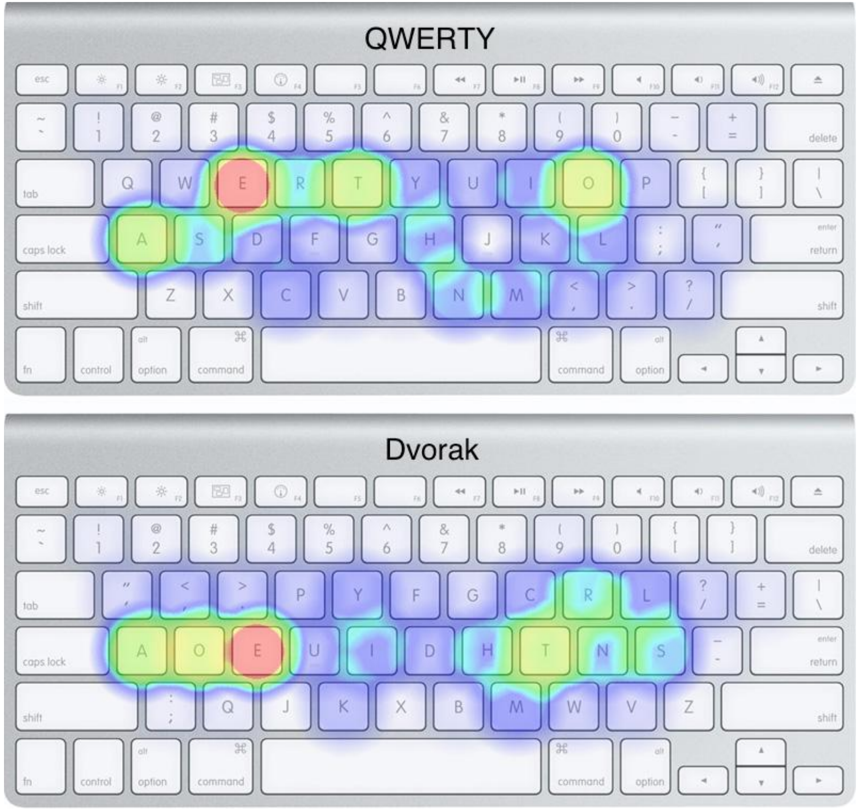

Á nícè sèntèncè wíth tòné màrks.
Á nícè sèntèncè wíth tòné màrks.

Hard to Use… (The Haptic & Psychological domains)

What does “hard to use” really mean?

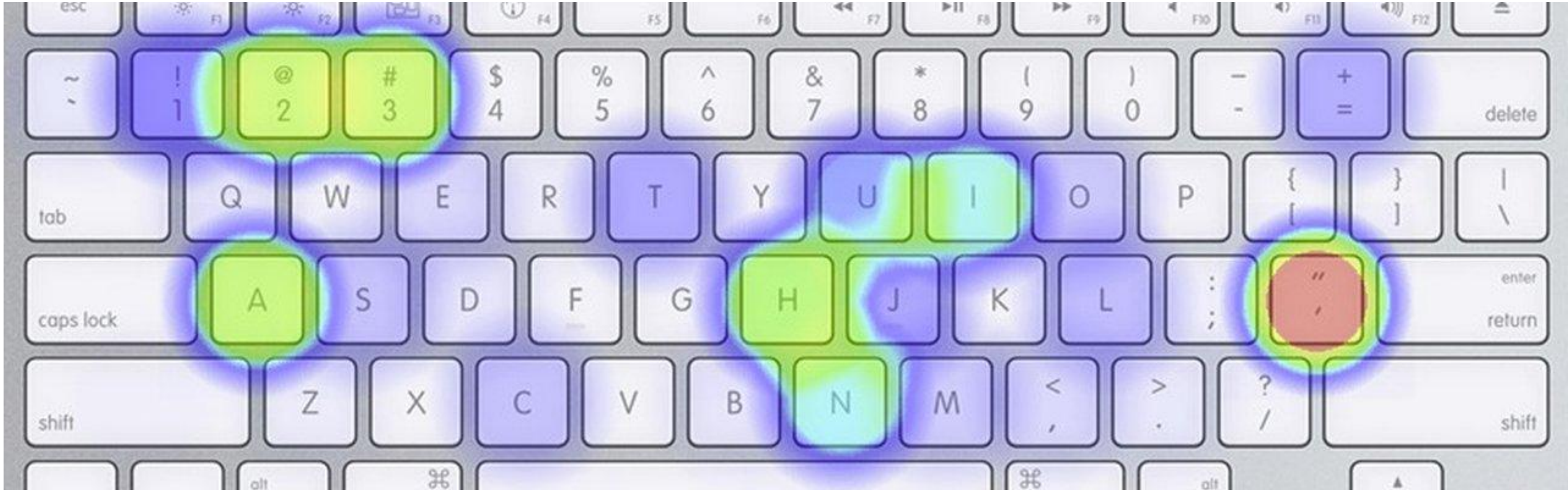
- QWERTY vs. Dvorak

“It is bonkers how much work this is. I just took 40 minutes to type 300 words and ran out of patience about 250 words ago. Getting a letter wrong makes me want to throw a chair. The S key being on the other side of the keyboard, under a pinky… I mean.” — CASEY JOHNSTON in *ars Technica*



Hard to Use… (the haptic domain)

What does “hard to use” really mean?


esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12
~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 – - + = delete
tab Q W E R T Y U I O P { [}] | \
caps lock A S D F G H J K L : ; " ' enter return
shift Z X C V B N M < , > . ? / shift
alt ⌘ ⌘ alt

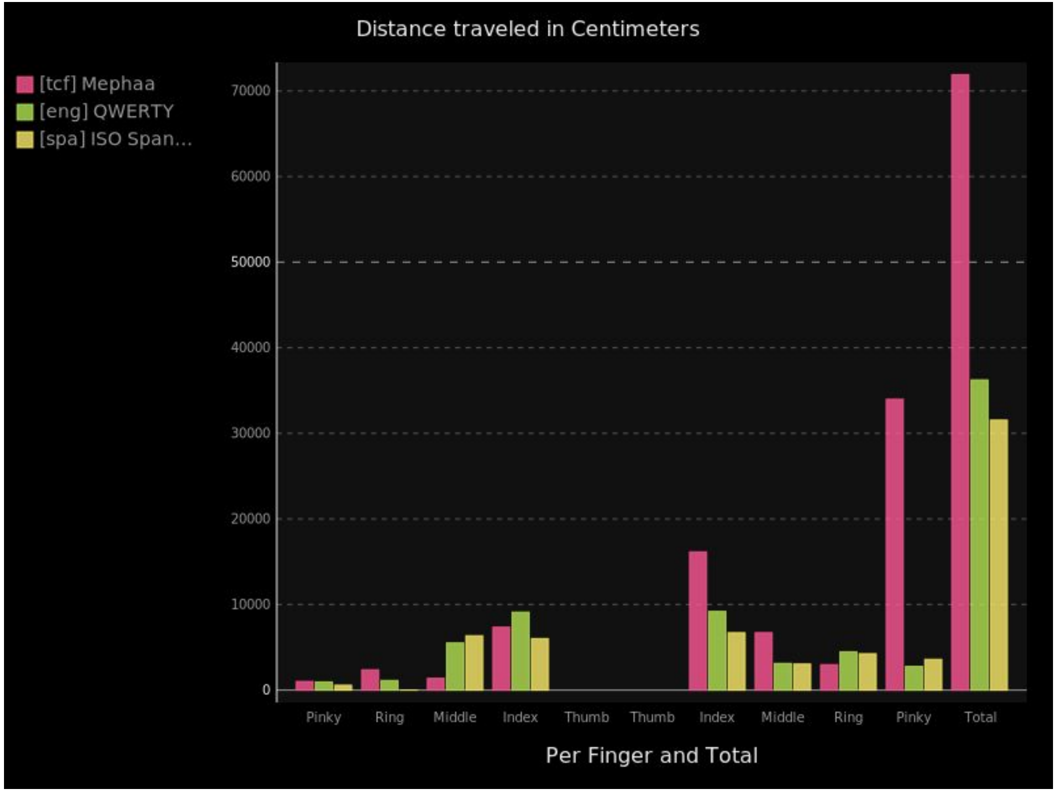
Hard to Use… (the distance measure)

What does “hard to use” really mean?

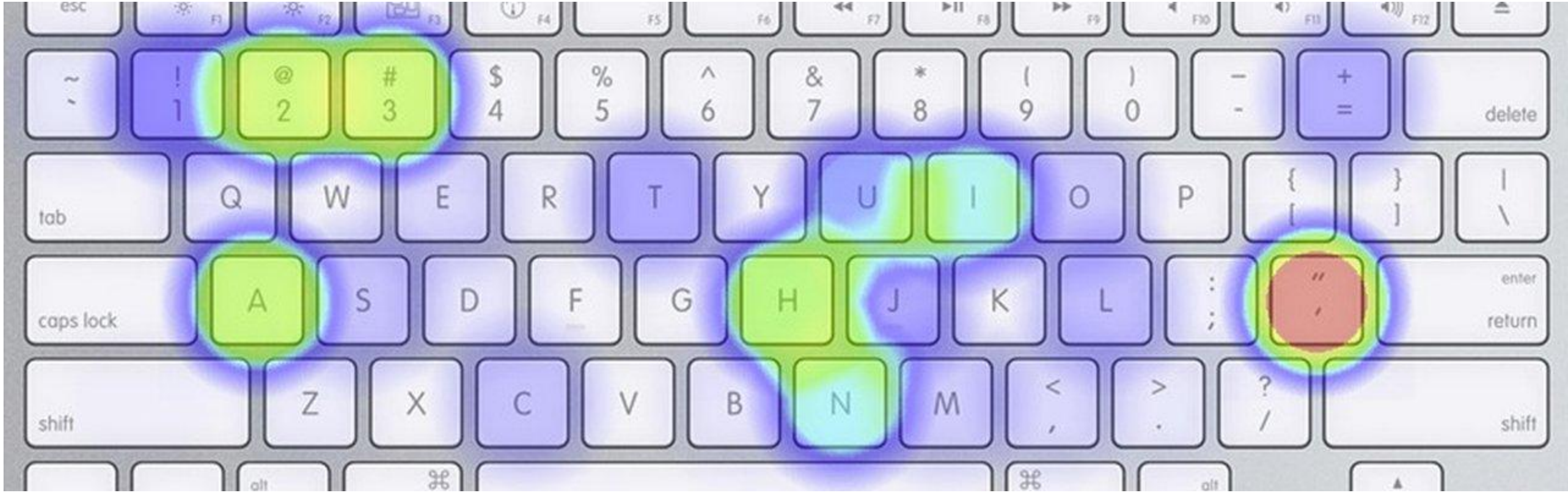
The same translated text across three languages:

- English - Green
- Spanish - Yellow
- Me'phaa - Rose

Right-hand heavy… Pinky heavy



Me'phaa text input patterns



Let’s As a Question:

Language as Instrument

Language as Object

Bourdieu, Pierre. 1977. “The economics of linguistic exchanges.” *Social Science Information* 16(6). 645–668. doi:10.1177/053901847701600601

Let’s As a Question:

What are the real cost factors when considering typing in multilingual communicative environments?... To what degree is the technology factor quantifiable?

Language as Instrument

Language as Object

Bourdieu, Pierre. 1977. “The economics of linguistic exchanges.” *Social Science Information* 16(6). 645–668. doi:10.1177/053901847701600601

The Experiment Design

- Use a Parallel corpus
- Select two keyboards per language
- Decompose the corpora to keystroke values
- Apply weight-based rankings to keystroke strings.
    - Weight-based rankings are applied on two factors:
        - location relative to rest position
        - Modeled finger
- Compare models without accounting for language

The Evidence Evaluated I

Keyboard Layout	Language
QWERTY	English
Dvorak	English
AZERTY	French
Bépo	French
AFU	Eastern Dan
Trans-Mande	Eastern Dan

The Evidence Evaluated II

James in English
(NLT)

James in French
(Darby 1885)

James in Eastern Dan
(Wycliffe 1991/2016)

Newspaper Collection
in Eastern Dan

The Evidence Evaluated III

Corpus	Orthographic Words	Unicode Grapheme Clusters	Unique Characters
eng-James	2,531	13,371	61
fra-James	2,378	13,372	68
dnj-James	5,197	23,958	61
dnj-Full	84,268	399,971	99

The Evidence Evaluated III

Keyboard Layout	Language	Fitness Score	Corpus
QWERTY	English	3362	James
Dvorak	English	1642	James
AZERTY	French	3358	James
Bépo	French	1472	James
AFU	French	3345	James
Trans-Mande	French	3588	James
AFU	Eastern Dan	17922	James
Trans-Mande	Eastern Dan	22723	James

The Evidence Evaluated III

For the Eastern Dan writer, it takes about **5.3 times the typing effort** when compared with the effort required to type the national language of the region.

The Economics of Linguistic Exchanges

For the Eastern Dan writer, it takes about **5.3 times the typing effort** when compared with the effort required to type the national language of the region.

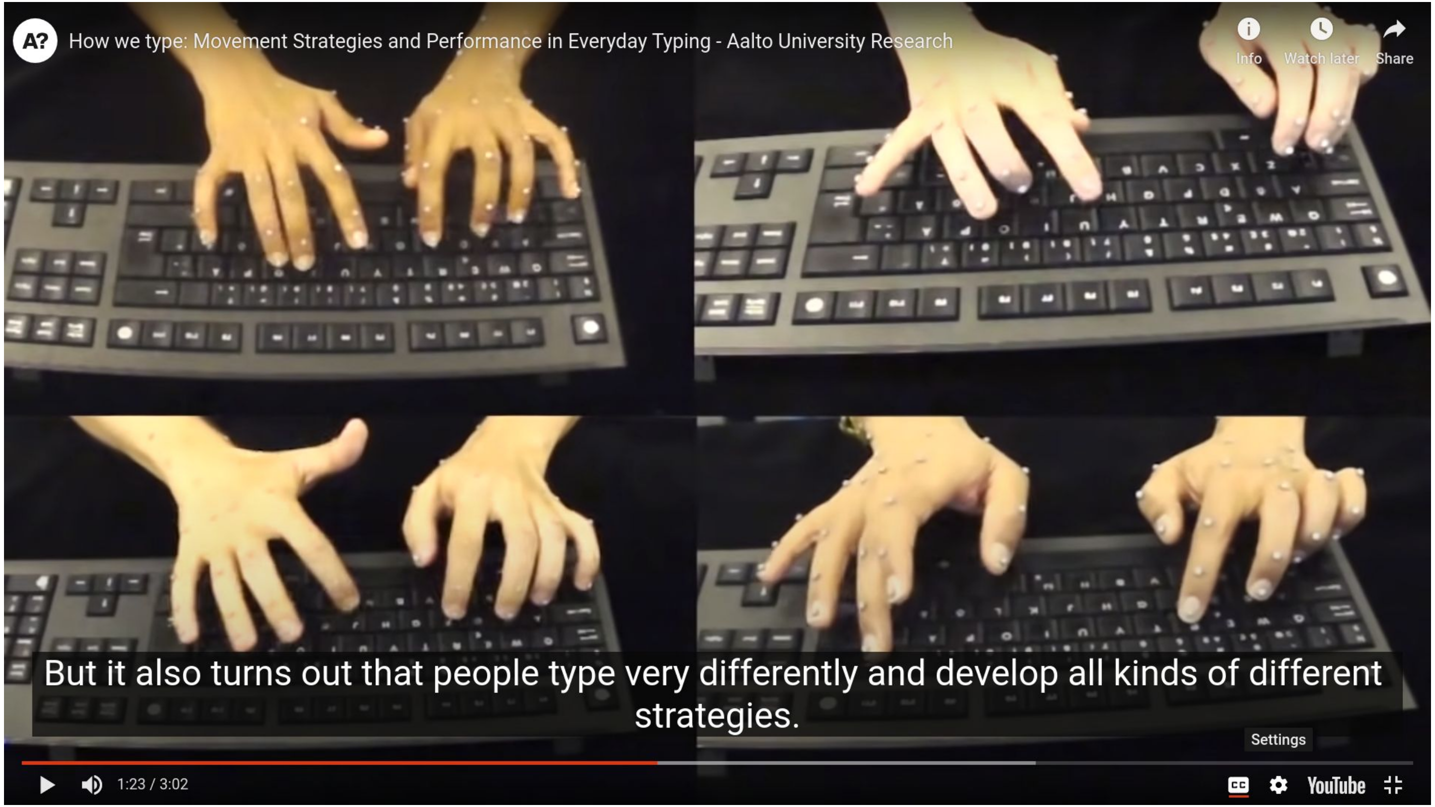
Language as Instrument

Language as Object

Multilingual environments allow for users to ask the questions: For what linguistic exchange is ___x___ language optimal?

Bourdieu, Pierre. 1977. “The economics of linguistic exchanges.” *Social Science Information* 16(6). 645–668. doi:10.1177/053901847701600601

The Evaluation Methods



How we type: Movement Strategies and Performance in Everyday Typing - Aalto University Research

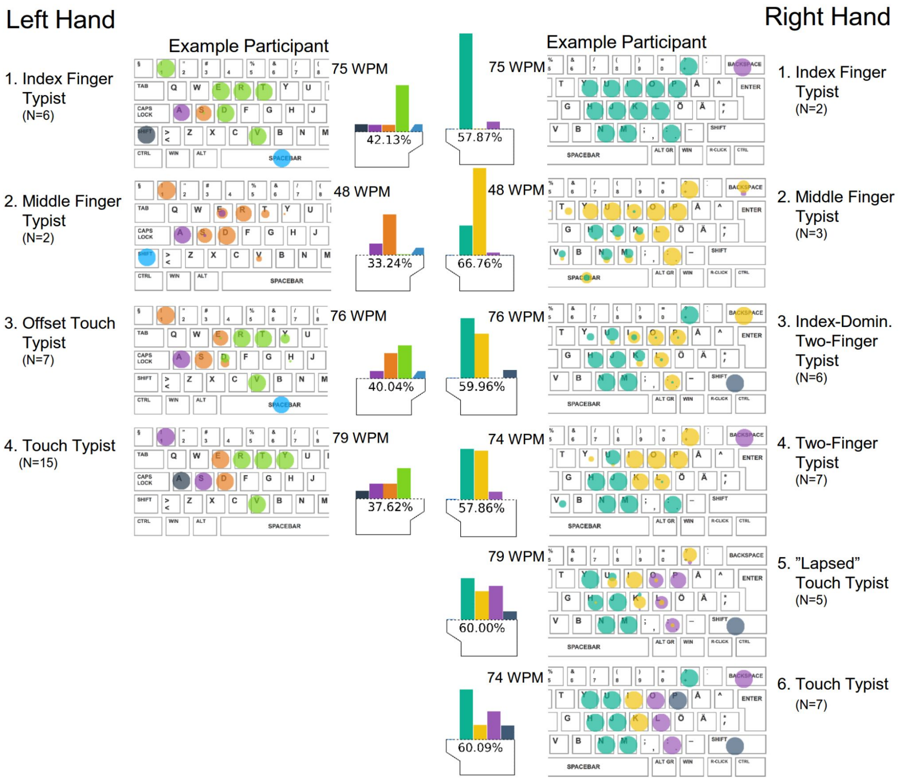
But it also turns out that people type very differently and develop all kinds of different strategies.

Evaluating the Standard Assumption

Feit et al. (2016) *How We Type: Movement Strategies and Performance in Everyday Typing*. — a must read for anyone working on typing.

- They present a typology of typing strategies
- Based on empirical observation
- Supported by open access data

- The Data includes the English and Finnish.
- They suggest that the standard model may not apply.



Future Experimental Work I

Currently the keyboard layout assignment problem is expressed as the following:

$$min \sum_{k} \sum_{l} Pkl \bullet Ckl$$

Where kl represent bigrams. And where C is a formulation of Shannon's law:

$$MT = a + b \log_2 \left(1 + \frac{D}{W}\right)$$

Future Experimental Work II

We have seen:

- not all bigrams (two unicode character sequences) are digraphs (two letter sequences representing a single orthographical unit).
- in latin script orthographies indicating tone, some digraphs are nonsequential (tone melodies/tone patterns).

Optimization to this point has focused on haptic interactions in languages where bigrams are generally represented by two “letters” and where multigraphs are sequential. Future work should integrate the tone orthography typology considerations, and the models known empirical models that typists use. Optimization should also look towards optimizing at the level of “least cognitively disruptive” rather than purely looking at character frequency.

Future Experimental Work III

- A more articulated algorithm for defining orthographic complexity in an effort to describe psychological reality.
  - Counting orthographic density by tonal melody/tonal pattern
  - Counting cognitively interruptive orthographic design, multi-graphs — both non-concatenative and concatenative
- Counting multi-stroke non-multigraphs
- Field testing and lab based testing with eye tracking, video observation, and motion sensing technology.

Cross-Disciplinary Relevant Questions

Psychology & Neuroscience:	*How does brain processing of haptic patterns align with brain processing of read patterns and spoken patterns in language?*
Sociology of Language:	***How interruptive is too interruptive? At which point will people choose a new language in which to communicate?***
Orthography and Writing Studies:	***How do pronunciation and meaning mappings to graphemes impact communication?***
Information and computational Science:	*How do we define the search-space and then how do we search it efficiently for optimal keyboard layout arrangements?*

Other Presentations and Publications

Publications:

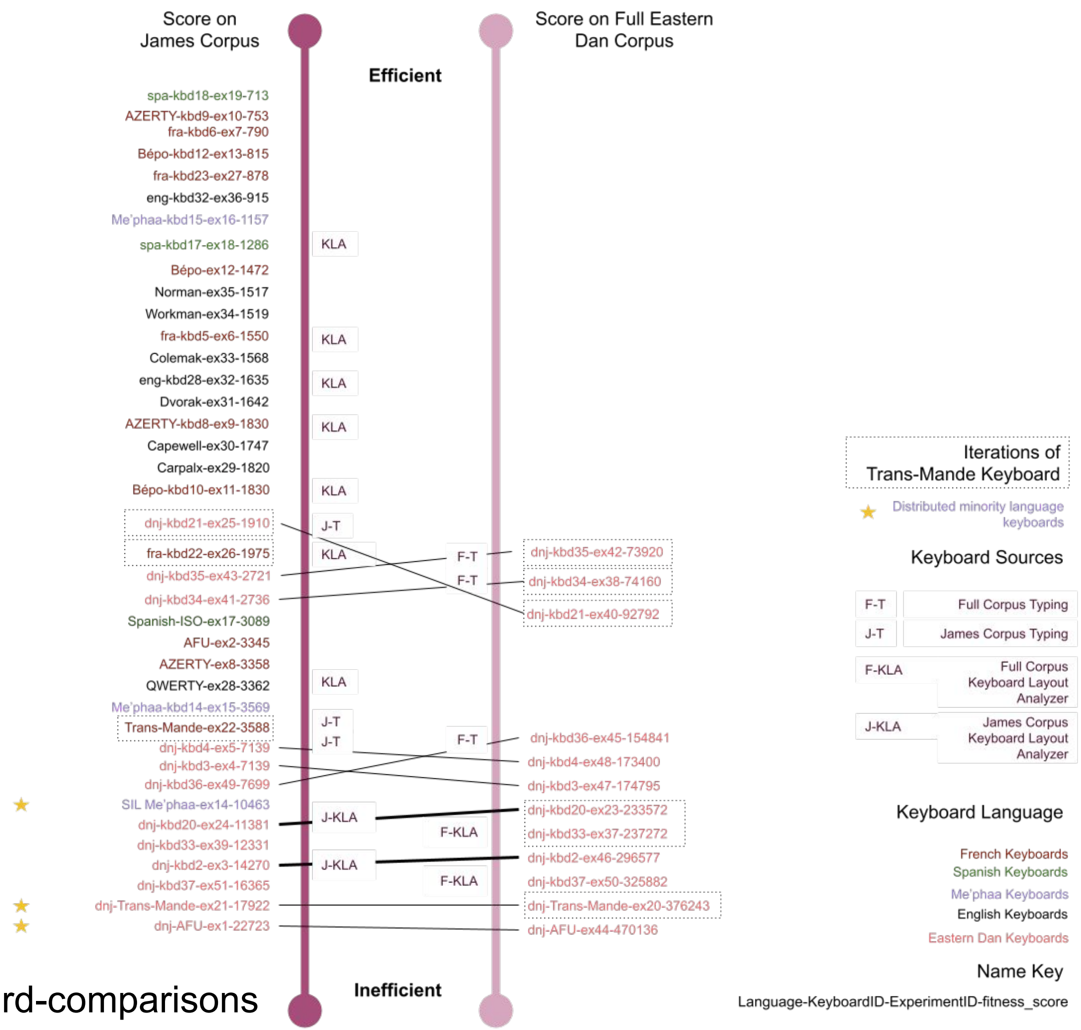
- Paterson III, Hugh J. 2015. **Keyboard Layouts: Lessons from the Me'phaa and Sochiapam Chinantec Designs**. In Mari C. Jones, Endangered Languages and New Technologies, 49–66. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781107279063.006

Presentations:

- Paterson III, Hugh J. 2015. **African Languages: Assessing the Text Input Difficulty.** Paper presented at: 46th Annual Conference of African Linguistics. University of Oregon. March 26th – 28th. https://hughandbecky.us/Hugh-CV/talk/2015-africa-assessing-thedifficulty-of-text-input
- Paterson III, Hugh J. 2015. **Assessing the Difficulty of the Text Input Task for Minority Languages**. Paper presented at: 4th International Conference on Language Documentation & Conservation. Ala Moana Hotel in Honolulu, HI. February 26th – March 1st. http://hdl.handle.net/10125/25318
- Paterson III, Hugh J. 2012. **Keyboard Layout as Part of Language Documentation:The Case of the Me'phaa and Chinantec Keyboards**. Paper presented at: Language Endangerment: Methodologies and New Challenges, CRASSH. Cambridge, United Kingdom. July 6th. https://hughandbecky.us/Hugh-CV/talk/2012-keyboard-layout-presentation

A Parallel Look

https://hughandbecky.us/Hugh-CV/post/keyboard-comparisons



Score on James Corpus

Score on Full Eastern Dan Corpus

Efficient

spa-kbd18-ex19-713
AZERTY-kbd9-ex10-753
fra-kbd6-ex7-790
Bépo-kbd12-ex13-815
fra-kbd23-ex27-878
eng-kbd32-ex36-915
Me'phaa-kbd15-ex16-1157
spa-kbd17-ex18-1286 KLA
Bépo-ex12-1472
Norman-ex35-1517
Workman-ex34-1519
fra-kbd5-ex6-1550 KLA
Colemak-ex33-1568
eng-kbd28-ex32-1635 KLA
Dvorak-ex31-1642
AZERTY-kbd8-ex9-1830 KLA
Capewell-ex30-1747
Carpalx-ex29-1820
Bépo-kbd10-ex11-1830 KLA
dnj-kbd21-ex25-1910 J-T
fra-kbd22-ex26-1975 KLA
dnj-kbd35-ex43-2721
dnj-kbd34-ex41-2736
Spanish-ISO-ex17-3089
AFU-ex2-3345
AZERTY-ex8-3358
QWERTY-ex28-3362 KLA
Me'phaa-kbd14-ex15-3569
Trans-Mande-ex22-3588 J-T
dnj-kbd4-ex5-7139 J-T
dnj-kbd3-ex4-7139
dnj-kbd36-ex49-7699
★ SIL Me'phaa-ex14-10463
dnj-kbd20-ex24-11381 J-KLA
dnj-kbd33-ex39-12331
dnj-kbd2-ex3-14270 J-KLA
dnj-kbd37-ex51-16365
★ dnj-Trans-Mande-ex21-17922
★ dnj-AFU-ex1-22723

F-T dnj-kbd35-ex42-73920
F-T dnj-kbd34-ex38-74160
dnj-kbd21-ex40-92792
F-T dnj-kbd36-ex45-154841
dnj-kbd4-ex48-173400
dnj-kbd3-ex47-174795
dnj-kbd20-ex23-233572
F-KLA dnj-kbd33-ex37-237272
dnj-kbd2-ex46-296577
F-KLA dnj-kbd37-ex50-325882
dnj-Trans-Mande-ex20-376243
dnj-AFU-ex44-470136

Inefficient

Iterations of Trans-Mande Keyboard

★ Distributed minority language keyboards

Keyboard Sources

F-T	Full Corpus Typing
J-T	James Corpus Typing
F-KLA	Full Corpus Keyboard Layout Analyzer
J-KLA	James Corpus Keyboard Layout Analyzer

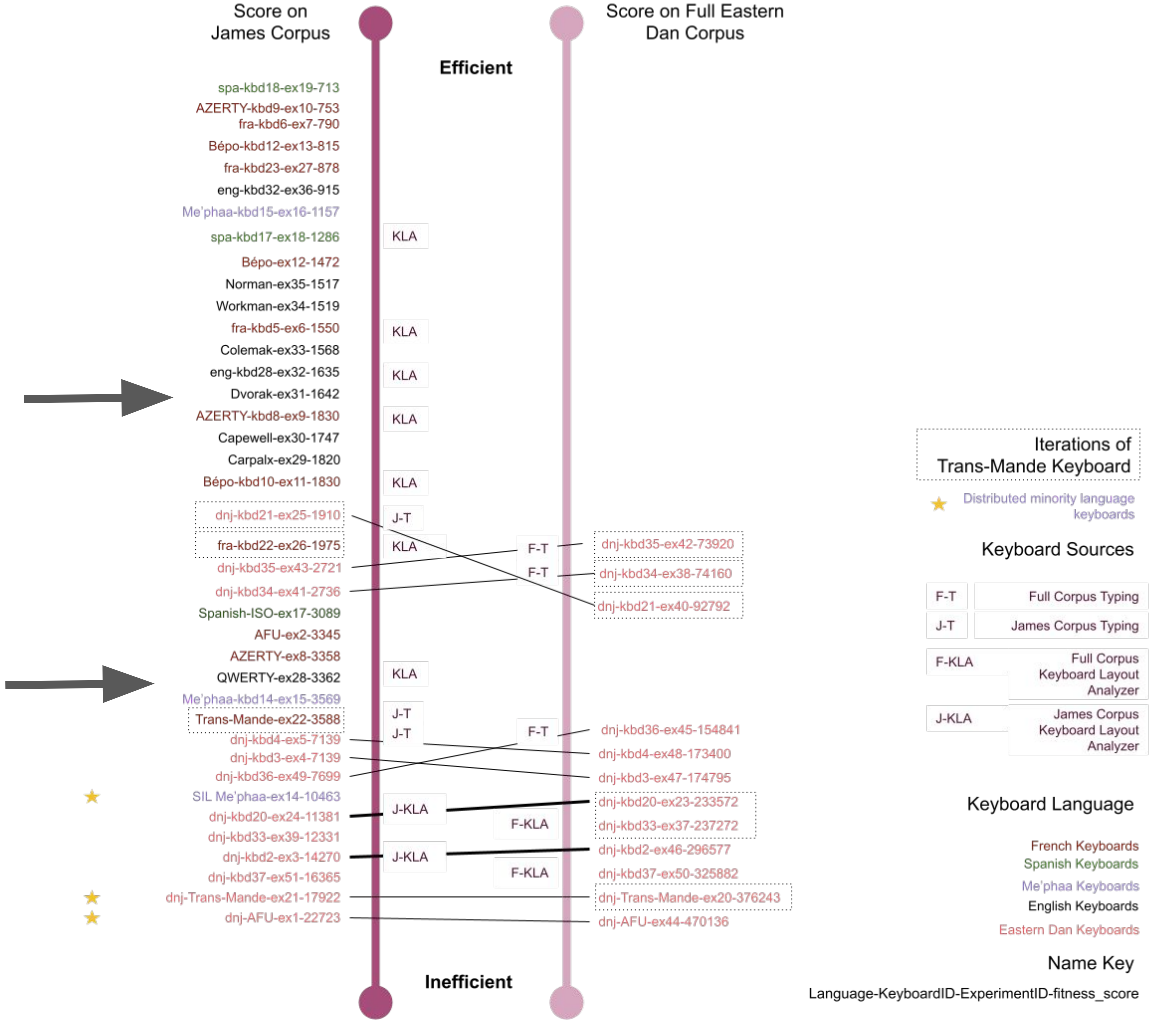
Keyboard Language

French Keyboards
Spanish Keyboards
Me'phaa Keyboards
English Keyboards
Eastern Dan Keyboards

Name Key

Language-KeyboardID-ExperimentID-fitness_score

A Parallel Look



Score on James Corpus

Score on Full Eastern Dan Corpus

Efficient

spa-kbd18-ex19-713
AZERTY-kbd9-ex10-753
fra-kbd6-ex7-790
Bépo-kbd12-ex13-815
fra-kbd23-ex27-878
eng-kbd32-ex36-915
Me'phaa-kbd15-ex16-1157
spa-kbd17-ex18-1286 KLA
Bépo-ex12-1472
Norman-ex35-1517
Workman-ex34-1519
fra-kbd5-ex6-1550 KLA
Colemak-ex33-1568
eng-kbd28-ex32-1635 KLA
Dvorak-ex31-1642
AZERTY-kbd8-ex9-1830 KLA
Capewell-ex30-1747
Carpalx-ex29-1820
Bépo-kbd10-ex11-1830 KLA
dnj-kbd21-ex25-1910 J-T
fra-kbd22-ex26-1975 KLA
dnj-kbd35-ex43-2721
dnj-kbd34-ex41-2736
Spanish-ISO-ex17-3089
AFU-ex2-3345
AZERTY-ex8-3358
QWERTY-ex28-3362 KLA
Me'phaa-kbd14-ex15-3569
Trans-Mande-ex22-3588 J-T
dnj-kbd4-ex5-7139 J-T
dnj-kbd3-ex4-7139
dnj-kbd36-ex49-7699
★ SIL Me'phaa-ex14-10463
dnj-kbd20-ex24-11381 J-KLA
dnj-kbd33-ex39-12331
dnj-kbd2-ex3-14270 J-KLA
dnj-kbd37-ex51-16365
★ dnj-Trans-Mande-ex21-17922
★ dnj-AFU-ex1-22723

F-T dnj-kbd35-ex42-73920
F-T dnj-kbd34-ex38-74160
dnj-kbd21-ex40-92792
F-T dnj-kbd36-ex45-154841
dnj-kbd4-ex48-173400
dnj-kbd3-ex47-174795
dnj-kbd20-ex23-233572
F-KLA dnj-kbd33-ex37-237272
dnj-kbd2-ex46-296577
F-KLA dnj-kbd37-ex50-325882
dnj-Trans-Mande-ex20-376243
dnj-AFU-ex44-470136

Inefficient

Iterations of Trans-Mande Keyboard

★ Distributed minority language keyboards

Keyboard Sources

F-T	Full Corpus Typing
J-T	James Corpus Typing
F-KLA	Full Corpus Keyboard Layout Analyzer
J-KLA	James Corpus Keyboard Layout Analyzer

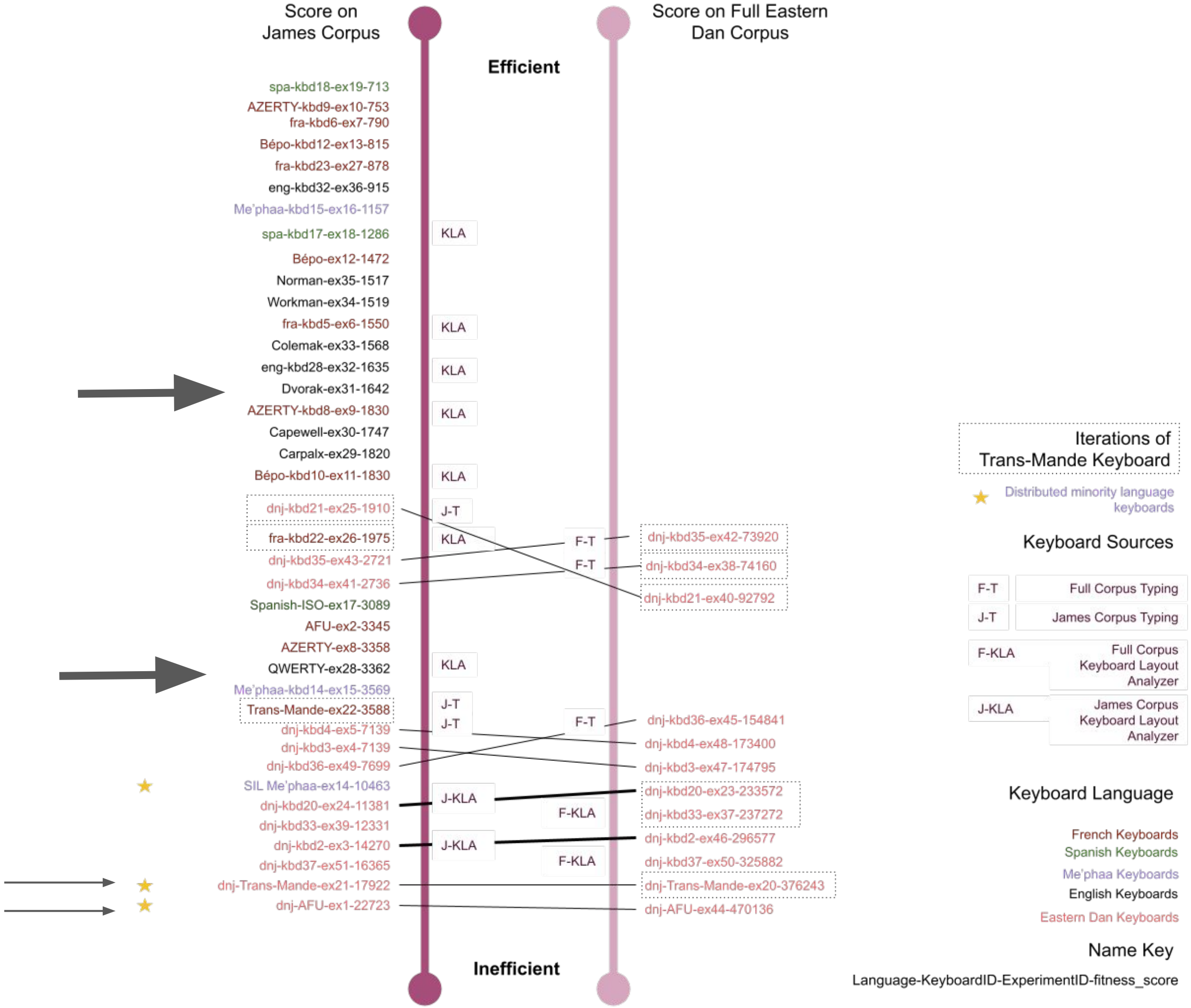
Keyboard Language

French Keyboards
Spanish Keyboards
Me'phaa Keyboards
English Keyboards
Eastern Dan Keyboards

Name Key

Language-KeyboardID-ExperimentID-fitness_score

A Parallel Look



Score on James Corpus

Score on Full Eastern Dan Corpus

Efficient

spa-kbd18-ex19-713
AZERTY-kbd9-ex10-753
fra-kbd6-ex7-790
Bépo-kbd12-ex13-815
fra-kbd23-ex27-878
eng-kbd32-ex36-915
Me'phaa-kbd15-ex16-1157
spa-kbd17-ex18-1286 KLA
Bépo-ex12-1472
Norman-ex35-1517
Workman-ex34-1519
fra-kbd5-ex6-1550 KLA
Colemak-ex33-1568
eng-kbd28-ex32-1635 KLA
Dvorak-ex31-1642
AZERTY-kbd8-ex9-1830 KLA
Capewell-ex30-1747
Carpalx-ex29-1820
Bépo-kbd10-ex11-1830 KLA
dnj-kbd21-ex25-1910 J-T
fra-kbd22-ex26-1975 KLA
dnj-kbd35-ex43-2721
dnj-kbd34-ex41-2736
Spanish-ISO-ex17-3089
AFU-ex2-3345
AZERTY-ex8-3358
QWERTY-ex28-3362 KLA
Me'phaa-kbd14-ex15-3569
Trans-Mande-ex22-3588 J-T
dnj-kbd4-ex5-7139 J-T
dnj-kbd3-ex4-7139
dnj-kbd36-ex49-7699
★ SIL Me'phaa-ex14-10463
dnj-kbd20-ex24-11381 J-KLA
dnj-kbd33-ex39-12331
dnj-kbd2-ex3-14270 J-KLA
dnj-kbd37-ex51-16365
★ dnj-Trans-Mande-ex21-17922
★ dnj-AFU-ex1-22723

F-T dnj-kbd35-ex42-73920
F-T dnj-kbd34-ex38-74160
dnj-kbd21-ex40-92792
F-T dnj-kbd36-ex45-154841
dnj-kbd4-ex48-173400
dnj-kbd3-ex47-174795
dnj-kbd20-ex23-233572
F-KLA dnj-kbd33-ex37-237272
dnj-kbd2-ex46-296577
F-KLA dnj-kbd37-ex50-325882
dnj-Trans-Mande-ex20-376243
dnj-AFU-ex44-470136

Inefficient

Iterations of Trans-Mande Keyboard

★ Distributed minority language keyboards

Keyboard Sources

F-T	Full Corpus Typing
J-T	James Corpus Typing
F-KLA	Full Corpus Keyboard Layout Analyzer
J-KLA	James Corpus Keyboard Layout Analyzer

Keyboard Language

French Keyboards
Spanish Keyboards
Me'phaa Keyboards
English Keyboards
Eastern Dan Keyboards

Name Key

Language-KeyboardID-ExperimentID-fitness_score