THESIS SEMINAR MEETING: MAY 3, 2021

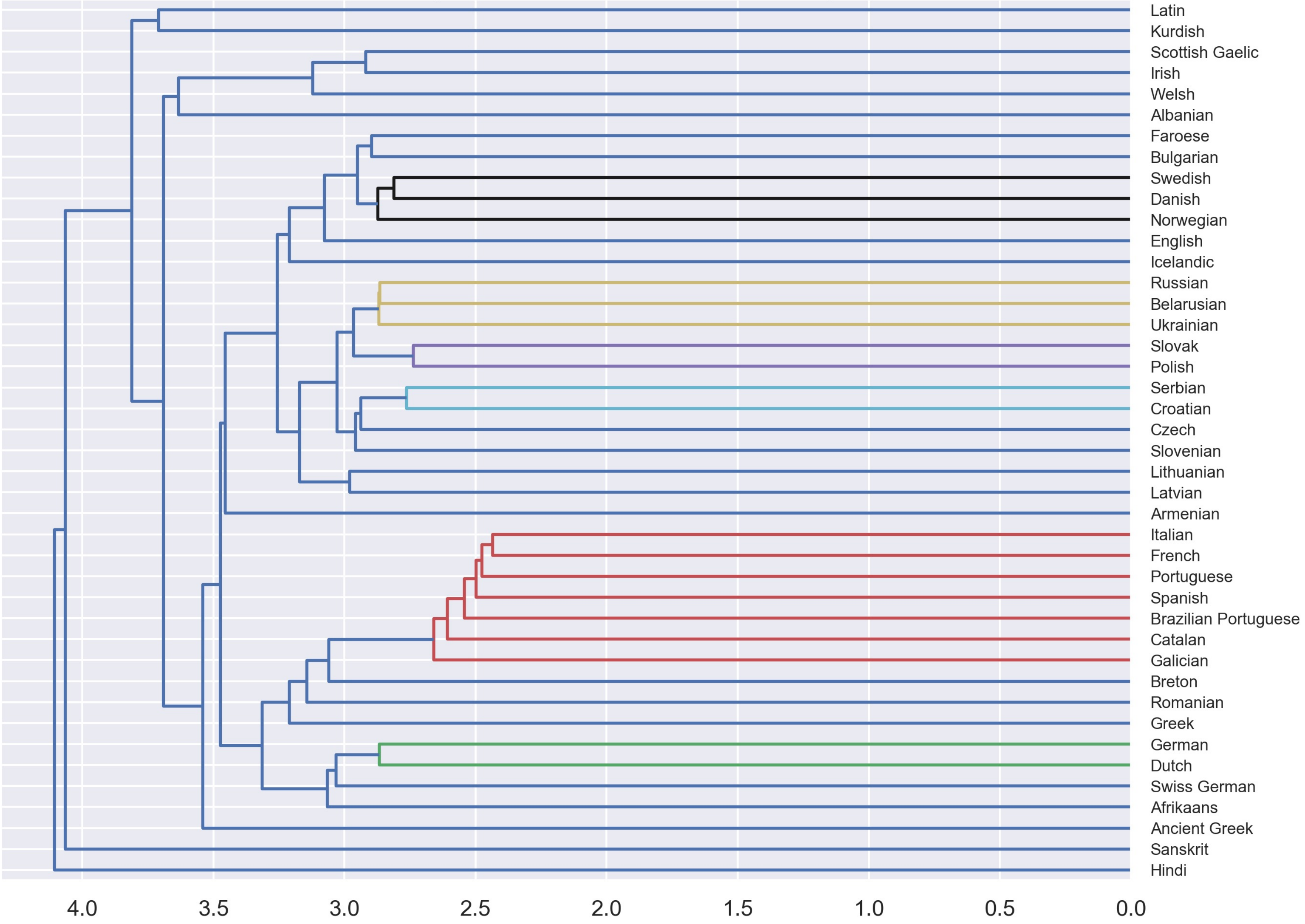
Philip Georgis

A QUICK TEST OF UNIVERSAL DEPENDENCIES

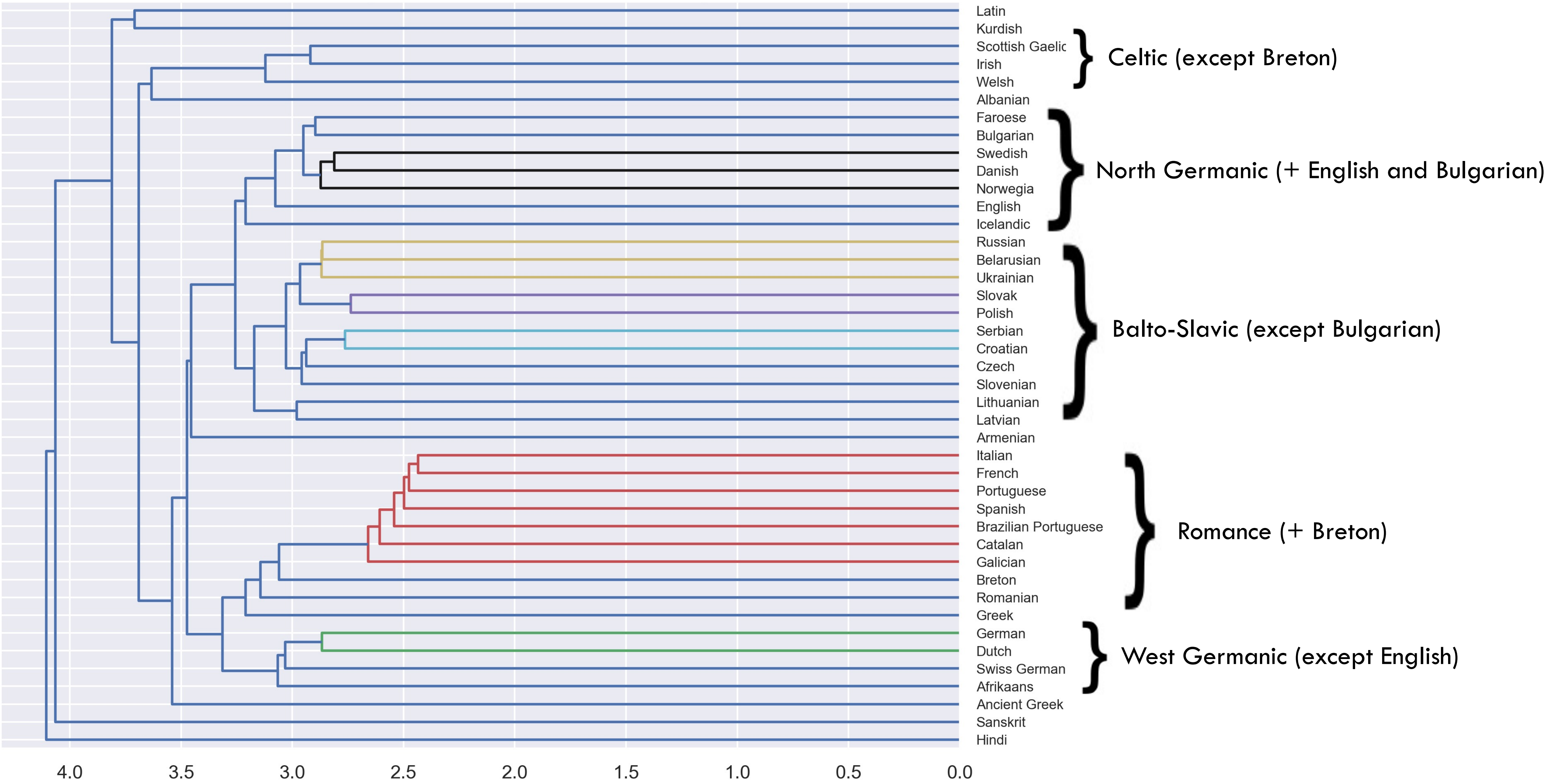
- Automatically processed UD corpora for 42 Indo-European languages
- Counted all trigrams of POS tags
- Calculated pairwise distance between languages as the mean surprisal of all POS trigrams (and then average both directions together)
  - e.g. POS_Surprisal(Dutch|English) = mean(surprisal of every Dutch POS trigram given English)
    POS_Surprisal(English|Dutch) = mean(surprisal of every English POS trigram given Dutch)
    Distance(Dutch, English) = mean(POS_Surprisal(Dutch|English), POS_Surprisal(English|Dutch))

- No preprocessing, normalization, or controls for corpus/sample size
- No lexical data available other than the POS tag sequences

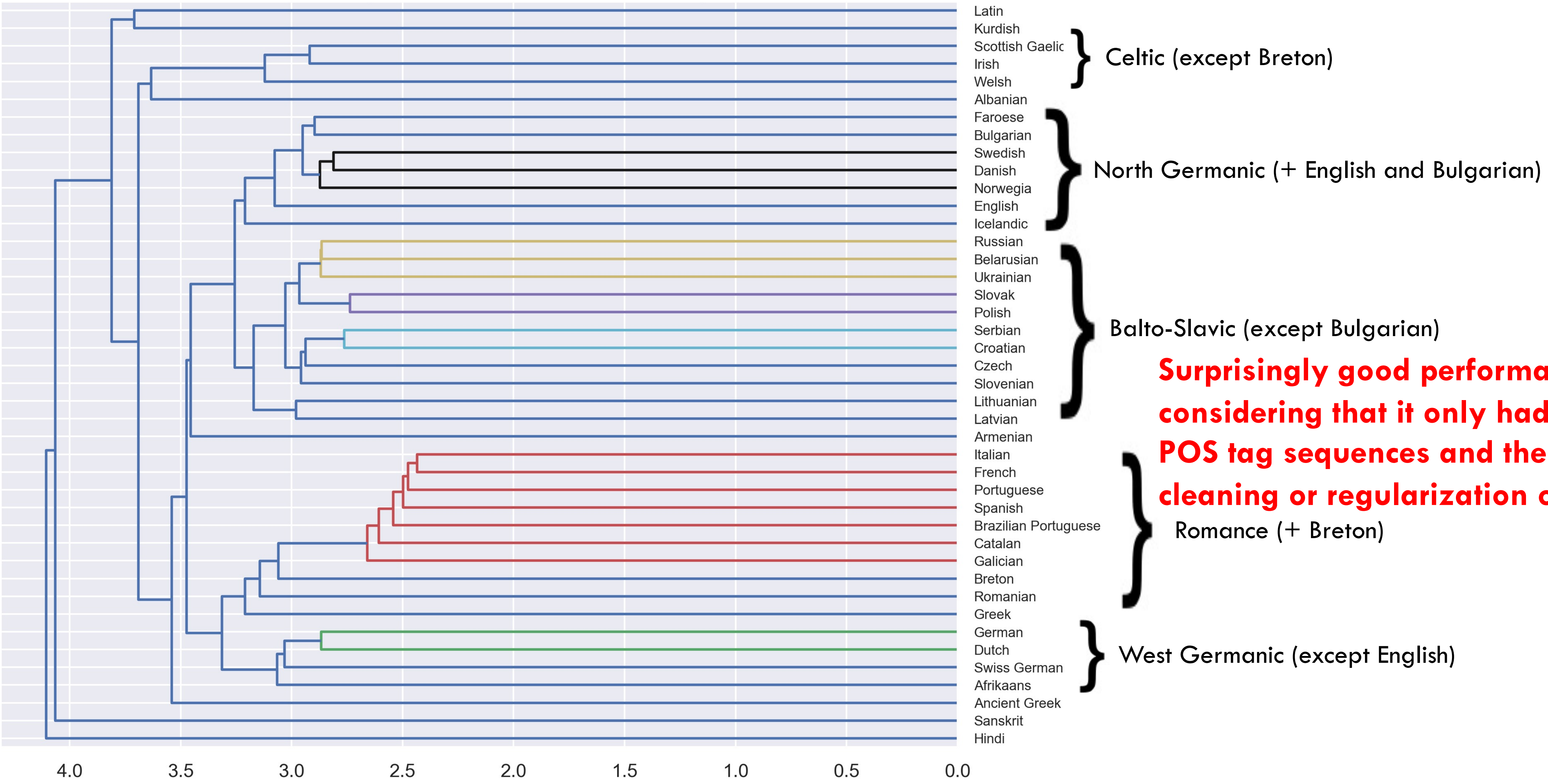
A QUICK TEST OF UNIVERSAL DEPENDENCIES



A QUICK TEST OF UNIVERSAL DEPENDENCIES



A QUICK TEST OF UNIVERSAL DEPENDENCIES



Surprisingly good performance, considering that it only had access to POS tag sequences and there was no cleaning or regularization of data!

UNIVERSAL DEPENDENCIES CORPORA

- **Benefits:** extensively annotated for morphology and syntax; easily machine readable

UNIVERSAL DEPENDENCIES CORPORA

- **Benefits:** extensively annotated for morphology and syntax; easily machine readable
- **Disadvantage:** not much in-depth coverage of families other than Indo-European

Indo-European:	59 languages
Uralic:	11 languages
Afro-Asiatic:	8 languages
Turkic:	4 languages
Sino-Tibetan:	3 languages
Niger-Congo:	2 languages
Austronesian:	2 languages
Dravidian:	2 languages
Other families:	21 languages combined

OTHER CORPORA
(E.G. BIBLE TRANSLATIONS, UNIVERSAL DECLARATION OF HUMAN RIGHTS)

- **Benefits:**
  - parallel according to verse/article
  - extensive coverage of hundreds of languages, not only European

- **Disadvantage:** no annotation!
  - How to extract syntactic parameters if there is no annotation?
  - Still possible to extract some data (e.g. Wälchli's (2019) study of feminine anaphors), but much more costly
  - Significantly narrows down range of possible parameters to explore
  - Reliance on "seed gram" (e.g. "she", "her", "hers" for Wälchli's study)

RESEARCH QUESTION DECISION

- **Option A:** Comparative Approach
  - Explore relative usefulness of lexical, phonetic, and morphosyntactic distances
  - Restrict to (Indo-)European languages due to lack of annotated corpus data for other groups

RESEARCH QUESTION DECISION

- **Option A:** Comparative Approach
  - Explore relative usefulness of lexical, phonetic, and morphosyntactic distances
  - Restrict to (Indo-)European languages due to lack of annotated corpus data for other groups
  - Similar study already done: Longobardi et al. (2016) - *Correlated Evolution or Not? Phylogenetic linguistics with syntactic, cognacy and phonetic data*
    - Only Indo-European languages
    - Binary syntactic parameters
    - Phonetic distances using ASJP sound classes and “Dice distance”

RESEARCH QUESTION DECISION

- **Option A:** Comparative Approach
  - Explore relative usefulness of lexical, phonetic, and morphosyntactic distances
  - Restrict to (Indo-)European languages due to lack of annotated corpus data for other groups
  - Similar study already done: Longobardi et al. (2016) - *Correlated Evolution or Not? Phylogenetic linguistics with syntactic, cognacy and phonetic data*
    - Only Indo-European languages
    - Binary syntactic parameters
    - Phonetic distances using ASJP sound classes and “Dice distance”

- **Option B:** In-Depth Lexico-Phonetic Approach
  - Study different methods of cognate evaluation
  - Possible sub-methods
    - Phonetic methods: sound classes, phonetic features, weighting schemes
    - Information theory methods: PMI, surprisal
  - No need to restrict to only Indo-European languages
    - Only need wordlists and phonetic data – much more widely available

RESEARCH QUESTION DECISION

- **Option A:** Comparative Approach
  - Explore relative usefulness of lexical, phonetic, and morphosyntactic distances
  - Restrict to (Indo-)European languages due to lack of annotated corpus data for other groups
  - Similar study already done: Longobardi et al. (2016) - *Correlated Evolution or Not? Phylogenetic linguistics with syntactic, cognacy and phonetic data*
    - Only Indo-European languages
    - Binary syntactic parameters
    - Phonetic distances using ASJP sound classes and "Dice distance"

- **Option B:** In-Depth Lexico-Phonetic Approach
  - Study different methods of cognate evaluation
  - Possible sub-methods
    - Phonetic methods: sound classes, phonetic features, weighting schemes
    - Information theory methods: PMI, surprisal
  - No need to restrict to only Indo-European languages
    - Only need wordlists and phonetic data – much more widely available

OPTION B: LEXICO-PHONETIC APPROACH

- Phonetic evaluation methods: measure how phonologically similar cognate pairs are

  - **Phonetic features:** measure phonetic distance between /d/ and /dʲ/, /ʐ/ and /rʲ/, etc. according to shared phonetic features
    - Simplest: normalized Levenshtein distance with substitution costs = phonetic distance between phone pairs
    - More complex: possibility for experimentation with weighting schemes for different segment types and/or penalties for deletions
    - Tonemes for Sino-Tibetan languages
    - Not specific to language pairs: /d/ to /dʲ/ would have the same distance in Polish-Russian as in Polish-Irish, etc.

e.g. Polish *drzewo* [dʐˈɛvɔ] and Russian дерево [dʲˈerʲɪvə] ‘tree’

d		**ʐ**	**ˈɛ**	**v**	**ɔ**
dʲ	**ˈe**	**rʲ**	**ɪ**	**v**	**ə**

OPTION B: LEXICO-PHONETIC APPROACH

- Phonetic evaluation methods: measure how phonologically similar cognate pairs are
  - **Sound classes:** measure distance of sound class of /ʐ/ to sound class of /rʲ/, etc.
    - Use ASJP sound classes
    - Two possibilities for measurement, either:
      - Based on phonetic features of sound class
        - but which phone within sound class to select as representative? e.g. ASJP class 'X' contains /χ/, /ʁ/, /ħ/, /ʕ/, all with different phonetic features
      - Based on global correspondence probabilities among sound classes using gold cognate sets
        - e.g. /t/ → /t͡ʃ/ is more likely than /t/ → /k/

e.g. Polish *drzewo* [dʐˈɛvɔ] and Russian дерево [dʲˈerʲɪvə] 'tree'

d		ʐ	ˈɛ	v	ɔ
dʲ	ˈe	rʲ	ɪ	v	ə

OPTION B: LEXICO-PHONETIC APPROACH

- Phonetic evaluation methods: measure how phonologically similar cognate pairs are
  - **Sound classes:** measure distance of sound class of /ʐ/ to sound class of /rʲ/, etc.
    - Use ASJP sound classes
    - Two possibilities for measurement, either:
      - Based on phonetic features of sound class
        - but which phone within sound class to select as representative? e.g. ASJP class 'X' contains /χ/, /ʁ/, /ħ/, /ʕ/, all with different phonetic features
      - Based on global correspondence probabilities among sound classes using gold cognate sets
        - e.g. /t/ → /t͡ʃ/ is more likely than /t/ → /k/

e.g. Polish *drzewo* [dʐˈɛvɔ] and Russian дерево [dʲˈerʲɪvə] 'tree'

d		ʐ	ˈɛ	v	ɔ
dʲ	ˈe	rʲ	ɪ	v	ə

OPTION B: LEXICO-PHONETIC APPROACH

- Information theoretic methods
  - Specific to each language pair, measure *consistency* of sound changes, irrespective of what those changes are
  - **PMI:** measure of how strongly PL /d/ correlates with RU /dʲ/, /ʐ/ with /rʲ/, etc.

$$i(x, y) := \log \frac{p(x, y)}{p(x)p(y)}$$

  - **Word Adaptation Surprisal:** measure of how *un*expected RU /dʲ/ is given PL /d/, etc.

$$WAS = \frac{1}{n} \sum_{i=1}^{n} -log_2 P(L1_i | L2_i)$$

e.g. Polish *drzewo* [dʐˈɛvɔ] and Russian дерево [dʲˈerʲɪvə] 'tree'

d		ʐ	ˈɛ	v	ɔ
dʲ	ˈe	rʲ	ɪ	v	ə

TREE EVALUATION METRICS

- “Genetic method” of Maurits & Griffiths (2014) [cited in Dediu, 2018]
  - Based on topology of gold standard tree
  - Uses only number of intermediate nodes/splits, branch lengths presumed to be unknown
  - Distance *d* between two related languages sharing *n* intermediate nodes on their path to the root is calculated as:

$$d = M - \sum_{i=1}^{n} \alpha^{i}$$

where *M* is the maximum distance and α is fixed at 0.69

TREE EVALUATION METRICS

- Dediu, 2018: *Making genealogical language classifications available for phylogenetic analysis*
  - **Problem:** no standardized trees for evaluation, branch lengths are not available from Ethnologue and Glottolog trees
  - Combines classification data/trees from Ethnologue and Glottolog with typological, lexical, and geographic distances to aggregate standardized trees *with branch lengths*
  - Available via GitHub and implemented in R
    - → potential source of gold standard measures for evaluation

MUTUAL INTELLIGIBILITY METRICS

- Two main types of experimental methods (Tang & van Heuven, 2015)
  - **Functional:** test informant's comprehension, count proportion of correctly translated words
  - **Opinion testing:** ask informant to rate their comprehension of the stimulus lect along a rating scale

- If mutual intelligibility is used as a basis for comparison, should probably ensure that the intelligibility measures are of the same type
  - Chinese dialects (Tang & van Heuven, 2015): both types
  - Slavic languages (SFB C4 studies): functional
  - Germanic, Romance, Slavic (Gooskens et al., 2018): functional [cloze tasks]
  - Arabic dialects (Čéplö et al., 2016): functional
  - Turkic (Lindsay): opinion testing

WORKS CITED

Čéplö, S., Bátora, J., Benkato, A., Milička, J., Pereira, C., & Zemánek, P. (2016). Mutual intelligibility of spoken Maltese, Libyan Arabic, and Tunisian Arabic functionally tested: A pilot study. *Folia Linguistica*, *50*(2), 583–628.

Dediu, D. (2018). Making genealogical language classifications available for phylogenetic analysis: Newick trees, unified identifiers, and branch length. *Language Dynamics and Change*, *8*(1), 1–21. https://doi.org/doi:10.1163/22105832-00801001

Gooskens, C., van Heuven, V. J., Golubović, J., Schüppert, A., Swarte, F., & Voigt, S. (2018). Mutual intelligibility between closely related languages in Europe. *International Journal of Multilingualism*, *15*(2), 169–193.

Longobardi, G., Buch, A., Ceolin, A., Ecay, A., Guardiano, C., Irimia, M., Michelioudakis, D., Radkevich, N., & Jäger, G. (2016). Correlated Evolution or Not? Phylogenetic linguistics with syntactic, cognacy and phonetic data. *Proceedings of the 11th International Conference (EVOLANG11)*.

Tang, C., & van Heuven, V. J. (2015). Predicting mutual intelligibility of Chinese dialects from multiple objective linguistic distance measures. *Linguistics*, *53*(2), 285–311.

Wälchli, B. (2019). The feminine anaphoric gender gram, incipient gender marking, maturity, and extracting anaphoric gender markers from parallel texts. In F. Di Garbo, B. Olsson, & B. Wälchli (Eds.), *Grammatical gender and linguistic complexity: Volume II: World-wide comparative studies* (Vol. 2, pp. 61–131). Language Science Press.