A Phoneme Clustering Algorithm Based on the Obligatory Contour Principle

University of Colorado **Boulder**

Mans Hulden
mans.hulden@colorado.edu
https://github.com/cvocp/cvocp

Hierarchical Clustering

Objective

Divide all phonemes/character types in a corpus into two sets S' and S'' such that an alternation-counting objective function is maximized.

This is motivated by the **Obligatory Contour Principle** in phonology which says that globally, "similarity" is avoided between adjactent segments (and tiers), particularly as regards **place of articulation**, and **tone**

Example corpus = **abracadabra**

abracadabra (bad split, 2 alternations)
abracadabra (better split, 6 alternations)
abracadabra (best split, 8 alternations)

Result: S′ = {a} S′′ = {b,c,d,r}

Top-level algorithm (Simulated annealing)

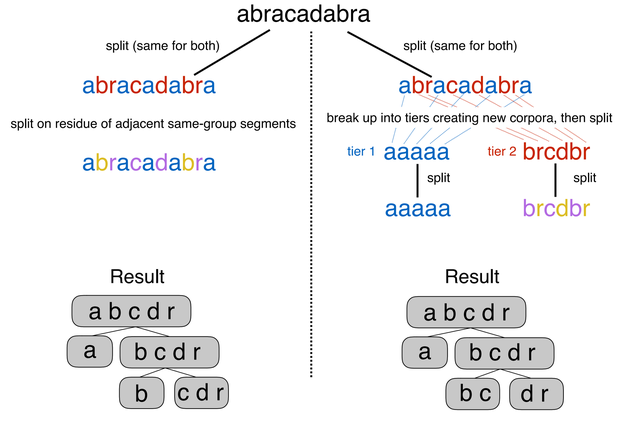
1. Randomly divide the set *S* into *S′* and *S′′*
2. Draw an integer *p* from Uniform(1 .. K), where K depends on a cooling schedule
3. Swap *p* random segments between S′ and S′′
4. If corpus score is higher after swap, keep swap else discard swap. Go to (2).

Recursion & Example

After the optimal top-level split is found as above, we can proceed resursively by either splitting on the **residue**, or dividing the corpus into two new subcorpora (**tiers**) and proceed. This gives us two variants of the main algorithm:

Two variants

residue-based recursion **tier**-based recursion



Phonemic Experiment (I)

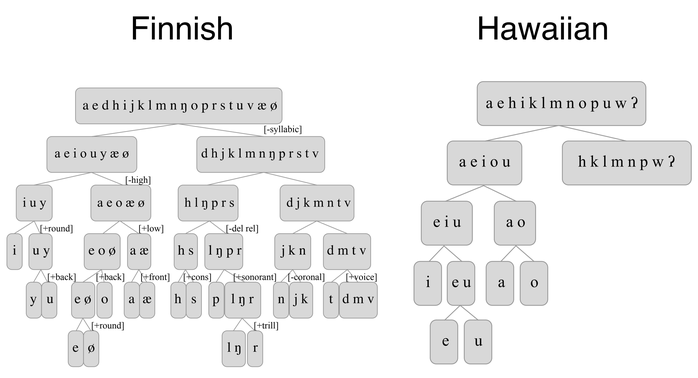
- Corpora from nine languages are featured (phonemic data)
- Measure how often resulting splits are describable by a single distinctive feature (both residue and tier-based methods)
- Separately see if top-level split is always consonants/vowels
- Compare against other algorithms for unsupervised discovery of C/V

Language	Source	Sample
Arapaho	(Cowell and Moss Sr, 2008)	towohei hiiθeti? tohnooke? toothei?eihoo . . .
Basque	Wikipedia + g2p	meʃikoko iriburuko espetʃe batean sartu zuten eta meʃiko . . .
English	(Brent and Cartwright, 1996)	ju want tu si ðə bʊk lʊk ðɛrz ə bɑɪ wɪð hɪz hæt . . .
Finnish	(Aho, 1884) + g2p	vai oli eilen kolmekymmentæ kotoapæinkø se matti ajelee . . .
Hawaiian	Wikipedia + g2p	ʔo ka ʔōlelo hawaiʔi ka ʔōlelo makuahine a ka poʔe maoli . . .
Hungarian	(Gervain and Erra, 2012)	idʒ nintʃ jɒj dɛ tʃɛtʃɛ hol ɒ montʃikɒ hol vɒn ɒ montʃi itt ɒ . . .
Italian	Wikipedia + g2p	tʃitta eterna kon abitanti e il komune piu popoloso ditalia . . .
Polish	(Boruta and Jastrzebska, 2012)	gdzie jest bartuc gdzie jest ɲe ma xodz tu a kuku tso xovaʃ . . .
Spanish	(Taulé et al., 2008) + g2p	un akuerdo entre la patronal i los sindikatos franθeses sobre . . .

Results (I)

Language	Splits OCP		Splits OCP(tier)		C/V (OCP)	C/V (Sukh.)	C/V (M&M)	Inventory size
Arapaho	9/14	(62.29)	11/15	(73.34)	100.0	100.0	100.0	16
Basque	8/14	(57.14)	16/20	(80.00)	100.0	100.0	100.0	21
English	3/12	(25.00)	15/25	(60.00)	**100.0**	21.62	94.59	37
Finnish	14/16	(87.50)	17/19	(89.47)	100.0	100.0	100.0	20
Hawaiian	4/5	(80.00)	8/12	(66.67)	**100.0**	**100.0**	92.30	13
Hungarian	10/20	(50.00)	21/31	(67.74)	**100.0**	96.97	**100.0**	33
Italian	7/11	(63.64)	15/20	(75.00)	100.0	100.0	100.0	22
Polish	10/21	(47.61)	23/33	(69.70)	**100.0**	**100.0**	97.30	37
Spanish	10/15	(66.67)	16/21	(76.19)	100.0	100.0	100.0	22

Example splits (I - residue method)



C/V distinctions (II)

- Evaluate ability to infer C/V (syllabic/non-syllabic) distinctions from graphemic data
- Data set from Kim & Snyder (2013): a Bible corpus in 503 languages
- Compare with other unsupervised algorithms:
  - Sukhotin (1962)
  - Moler & Morrison (1983)
  - Kim & Snyder (2013)
- Learn distinctions:
  - Individually (one language at a time)
  - All together (as one big corpus)
- Accuracy:
  - per token (for comparison w/ K&S)
  - per type

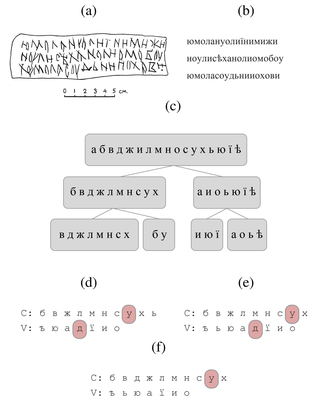
Results (II)

		OCP	Sukhotin	M&M	K&S
Individual	Type	**95.10**	92.50	94.15	–
	Token	**96.55**	93.65	95.59	95.99
All	Type	**96.43**	**96.43**	89.79	–
	Token	**99.89**	**99.89**	99.79	98.55

- Actual accuracy even higher due to 5 errors in gold

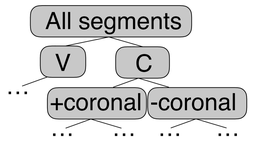
Short manuscripts

- Experiment example with extremely short manuscripts
- "Birch Bark Letter" no 292 (a), transcribed (b)
  - Oldest known text in Finnish languages (13th C.)
  - 54 letters
  - Contains variable spellings of the same word
  - Splits in (c) [residue method]
  - C/V in (f); (d) = M&M; (e) = Sukhotin
    - Errors marked in red



Tier-based algorithm & coronals

- Tier-based algorithm is very robust in splitting along [+coronal]/[-coronal] in second split:



- Results on graphemic data from 14 languages from Universal Dependencies corpora 2.0 (Nivre et al., 2017), with hypothesized +coronal split shown:

Language	Second Consonant Group	#C
Basque	(c) l n (ñ) r s x z	21
Catalan	l n r s x z	22
Irish	d l n r s	13
Dutch	h l n r x z	19
Estonian	h l n r s	16
Finnish	h l n r s (š) (x) (z)	21
German	j l n r s x z	21
Indonesian	l n r s z	20
Italian	h l n r s (y)	21
Latin	d h l n r s	16
Latvian	č j ķ l ļ n ņ r s z ž	24
Lithuanian	j l n r s š z ž	19
Portuguese	ç j l n (ñ) r s x	24
Slovak	c ď j l ĺ ľ n ň r s š z ž	26

Wrap-up

- The OCP seems to "hold" for syllabic/non-syllabic and coronal/non-coronal place of articulation, and frontness/backness of vowels
- Remaining splits are not robust along distinctive feature lines
- Algorithm is very good at detecting consonant/vowel (syllabic/non-syllabic) distinctions; better than previous efforts on all data sets
- Tier-based variant of algorithm is more robust and detects coronals with high accuracy

References

Young-Bum Kim and Benjamin Snyder. 2013. Unsupervised consonant-vowel prediction over hundreds of languages. In *Proceedings of ACL*. Sofia, Bulgaria, pages 1527–1536.

Cleve Moler and Donald Morrison. 1983. Singular value analysis of cryptograms. *American Mathematical Monthly* pages 78–87.

Boris V. Sukhotin. 1962. Eksperimental'noe vydelenie klassov bukv s pomoshch'ju EVM. *Problemy strukturnoj lingvistiki* pages 198–206.