Department of Statistics, The University of British Columbia

STAT 406 - Winter 2018 - Lecture # 21

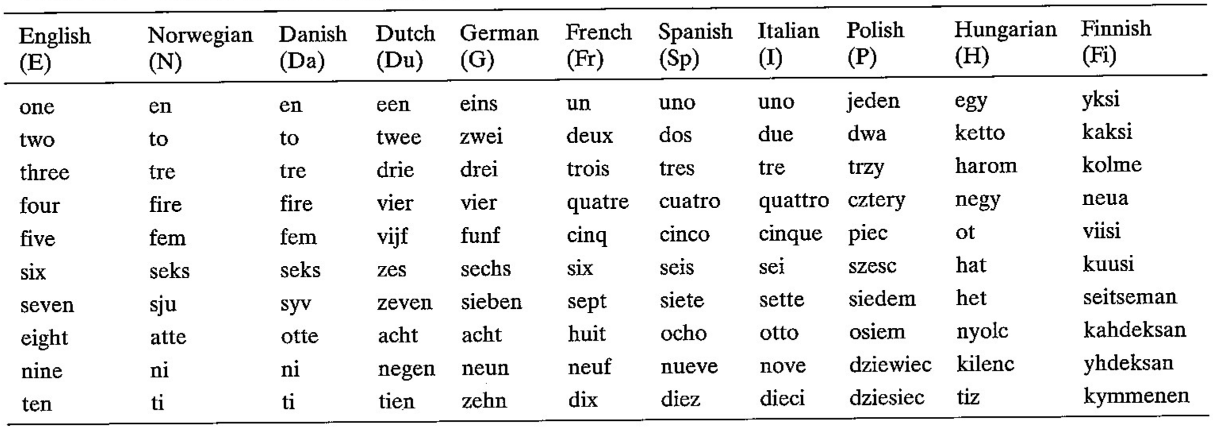
Matías Salibián Barrera

27 November 2018

1 In-class Activity

Consider the `numerals` in 11 European languages:

TABLE 12.3 NUMERALS IN 11 LANGUAGES



English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neua
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

We measure the dissimilarity between two languages as the number of numerals that start with a different letter. For example, eight numerals in French and Spanish start with the same letter (only "quatre"/ "cuatro" and "huit" / "ocho" differ). So $d(\text{Fr}, \text{Sp}) = 2$. The table below contains these dissimilarities for all possible pairs of languages:

	E	N	Da	Du	G	Fr	S	I	P	H	Fi
E											
N	2										
Da	2	1									
Du	7	5	6								
G	6	4	5	5							
Fr	6	6	6	9	7						
S	6	6	5	9	7	2					
I	6	6	5	9	7	1	1				
P	7	7	6	10	8	5	3	4			
H	9	8	8	8	9	10	10	10	10		
Fi	9	9	9	9	9	9	9	9	9	8	

Perform 2 or 3 iterations of an agglomerative hierarchical clustering algorithm using the single and average linkage criteria.
Recall that:

- Single linkage: the dissimilarity between clusters $\mathscr{C}_1$ and $\mathscr{C}_2$ is the smallest dissimilarity among any pair of elements:

$$d\left(\mathscr{C}_1, \mathscr{C}_2\right) = \min\left\{d(a_i, b_j),\ a_i \in \mathscr{C}_1,\ b_j \in \mathscr{C}_2\right\}$$

- Average linkage: the dissimilarity between clusters $\mathscr{C}_1$ and $\mathscr{C}_2$ is the average dissimilarity among all possible pairs of elements:

$$d\left(\mathscr{C}_1, \mathscr{C}_2\right) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} d(a_i, b_j)$$

  where $\mathscr{C}_1 = \{a_1, a_2, \dots, a_n\}$ and $\mathscr{C}_2 = \{b_1, b_2, \dots, b_m\}$