Orthographic differences among Germanic languages:

Stem variation versus inflectional affix variation

Wilbert Heeringa Femke Swarte Anja Schüppert
Charlotte Gooskens

University of Groningen
Faculty of Arts, Scandinavian Languages and Cultures

Workshop on Quantitative Linguistics and Dialectology
University of Groningen, June 29, 2012

Introduction

- Sometimes we may be confronted with texts in an unknown but closely-related language.
- To what extent are we able to understand the texts? This depends on the number of cognates, differences in orthography and syntax, etc.
- This study takes place in the context of a larger research program which aims to find the (non-)linguistic determinants of mutual intelligibility within the Germanic, Romance and Slavic language groups.
- Intelligibility scores of written and spoken language are obtained with a large-scale web-based experiment.
- What factors are predictors of these intelligibility scores?

Introduction

- Today we focus on orthography which is likely one of the most important predictors of the scores (of written language) obtained with the web-based experiment.
- Orthographic differences are the result of differences in:
  - writing system: e.g. Dutch *sector* versus German *Sektor*;
  - pronunciation: e.g. Dutch *helpen* versus German *helfen*.

Introduction

- We distinguish between differences in:
  - stems:
    - roots, compounds or derivational complexes
    - e.g. Dutch *helpen* versus German *helfen*
  - inflectional affixes:
    - in our data usually a suffix, in a few cases a prefix.
    - e.g. Dutch *regels* versus German *Regeln*
    - e.g. Dutch *gezien* versus English *seen*

Introduction

- The stem of a word is generally considered to have a larger information loading than its affix.
- Are stems more strongly affected by language change and therefore do they show more diversity?
- If so, orthographic stem variation may be stronger than orthographic affix variation, and the two types of variation may be not (strongly) correlated.

Hypotheses

- We hypothesize:
  1. Orthographic stem variation among languages does not correlate with orthographic variation in inflectional affixes.
  2. Orthographic stem variation among languages is larger than orthographic variation in inflectional affixes.
- We focus on Germanic languages:
  Danish, Dutch, English, German, Swedish.

Testing the hypotheses

- We test the hypotheses:
  - per language group:
    aggregated stem differences among the five languages are compared to the corresponding aggregated affix distances
  - per language pair:
    individual stem distances of the word pairs are compared to the corresponding individual affix distances for each of the ten language pairs.

Previous research in morphology

- Jean Séguy (1973). He used the *Atlas linguistique de la Gascogne* which includes 68 morpho-syntactic variables and 44 variables in verb morphology. 154 locations.
- Hans Goebl (1982, 1984, 1993). He used the *l'Atlas Linguistique de l'Italie et de la Suisse Méridionale*. 127 morpho-syntactic variables, 251 locations.
- Heeringa, Wieling, van den Berg, Nerbonne. They used the *Morphological Atlas of the Dutch Dialects*. 52 morphological variables. 130 sample sites in Low Saxony (Northeast of the Netherlands).
- All of these studies measure morphological variation on the basis of **categorical** variables.

Previous research in morphology

- Categorical variables may be *historically* determined, for example:
  - Is the plural suffix in *huize* 'houses' a realization of plural suffix *en* or *er*?
  - Is the prefix in the past participle *ewerkt* 'worked' a realization of prefix *ge* or a separate category?
- Our focus is not on history, but rather on intelligibility. Therefore we do not distinguish categories. Variation in both stems and affixes is measured with Levenshtein distance.
- Focussing on inflectional affix variation does not include all kinds of morphological variation.

Levenshtein distance

- Orthographic distances between stems and between affixes are measured with Levenshtein distance.
- Calculate the cost of changing one string into another.
- Example: English *interest* versus Swedish *intresse*:

English:	i	n	t	e	r	e	s	t	
Swedish:	i	n	t		r	e	s	s	e
				1				1	1

  A total cost of 3 divided by a length of 9 gives a word distance of 0.33 or 33%.
- Many sequence operations map *interest* $\longrightarrow$ *intresse*. Levenshtein distance = cost of cheapest mapping.

Weighing differences

- Two characters may be different because of differences in
  - the base:
    e.g. a vs. e, p vs. b, etc. We weigh this as 1.0
  - the diacritic:
    e.g. a vs. á, à vs. á, etc. We weigh this as 0.3.
- Insertions and deletions are weighed as 1.0.

Alignment

- We assure that the minimum cost is based on a alignment in which:
  - a vowel matches with a vowel
  - a consonant matches with a consonant

Small example Dutch versus German

	Dutch	German	number of different characters	total number of characters per word	proportion of different characters
1	helpen	helfen	1	6	0.17
2	monden	Münder	2	6	0.33
3	regels	Regeln	1	6	0.17
4	bakken	backen	1	6	0.17
5	gezegd	gesagt	3	6	0.50
					0.27

The average distance is the aggregated distance which is 0.27 or 27%.

Small example Dutch versus German: stem distances

	Dutch	German	number of different characters	total number of characters per word	proportion of different characters
1	**help**+en	**helf**+en	1	4	0.25
2	**mond**+en	**Münd**+er	1	4	0.25
3	**regel**+s	**Regel**+n	0	5	0.00
4	**bakk**+en	**back**+en	1	4	0.25
5	ge+**zeg**+d	ge+**sag**+t	2	3	0.67
					0.28

The average distance is the aggregated distance which is 0.28 or 28%.

Small example Dutch versus German: affix distances

	Dutch	German	number of different characters	total number of characters per word	proportion of different characters
1	help+**en**	helf+**en**	0	2	0
2	mond+**en**	Münd+**er**	1	2	0.50
3	regel+**s**	Regel+**n**	1	1	1
4	bakk+**en**	back+**en**	0	2	0
5	**ge**+zeg+**d**	**ge**+sag+**t**	1	3	0.33
					0.37

The average distance is the aggregated distance which is 0.37 or 37%.

The data set

- Translations of four English texts in each of the other four languages:

Text	Number of words
Child Athletes	236
Catching a cold	222
Driving in Winter	226
Riding a Bike	235
	919

- The translations are aligned to each other in columns.

The data set

- Since the English text cannot always be translated literally in each language, some words are missing and some words are added, causing empty cells in the table.

The data set: first part of the table

English	Danish	Dutch	German	Swedish
Parents	Forældre	Ouders	Eltern	Föräldrar
whose	hvis	wiens	deren	vars
children	born	kinderen	Kinder	barn
show	viser	tonen		
			haben	
a	en		ein	
special	særlig	speciale	besonderes	särskilt
interest	interesse	belangstelling	Interesse	idrottsintresserade
in	inden		an	
for		voor		
a	en	een	einer	
particular	bestemt	bepaalde	bestimmten	
sport	sportsgren	sport	Sportart	

The data set

- Number of word pairs per language pair:

	Danish	Dutch	English	German	Swedish
Danish		733	735	623	531
Dutch			759	650	529
English				630	515
German					475
Swedish					

The data set

- When listening to a foreign languages, affixes will be related to the corresponding affixes in the native language when the stems to which they belong to are cognates.
- The stem mainly carries the meaning of a word.
- Therefore orthographic distances are calculated for word pairs the members of which are cognates.

The data set

- Number and percentage of cognate pairs per language:

	Danish	Dutch	English	German	Swedish
Danish		288	276	262	390
Dutch	39%		384	411	195
English	38%	51%		283	150
German	42%	63%	45%		180
Swedish	73%	37%	29%	38%	

- The orthographic distance of a language pair is the sum of the cognate pair distances divided by the number of cognate pairs.

Stem distances

- Orthographic stem distances between languages in percentages:

	Danish	Dutch	English	German	Swedish
Danish		44	51	48	24
Dutch			53	45	47
English				59	55
German					48
Swedish					

Affix distances

- Orthographic affix distances between languages in percentages:

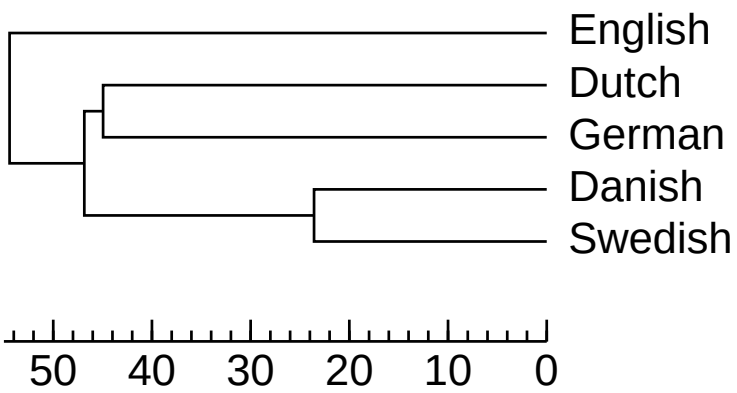
	Danish	Dutch	English	German	Swedish
Danish		32	29	35	26
Dutch			21	18	43
English				30	32
German					45
Swedish					

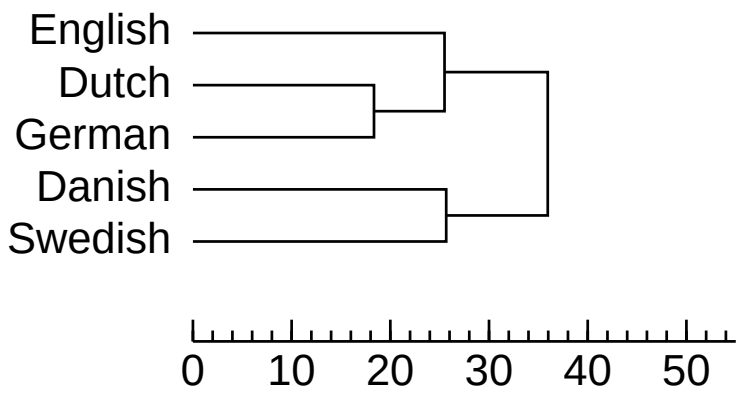
Cluster analysis

- We applied hierarchical cluster analysis to both the ‘stem’ and the ‘affix’ distances.
- Result is a binary tree structure (one for the ‘stem’ distances and another one for the ‘affix’ distances) in which the varieties are the leaves and the branches reflect the distances between the leaves, known as a *dendrogram*.

Cluster analysis

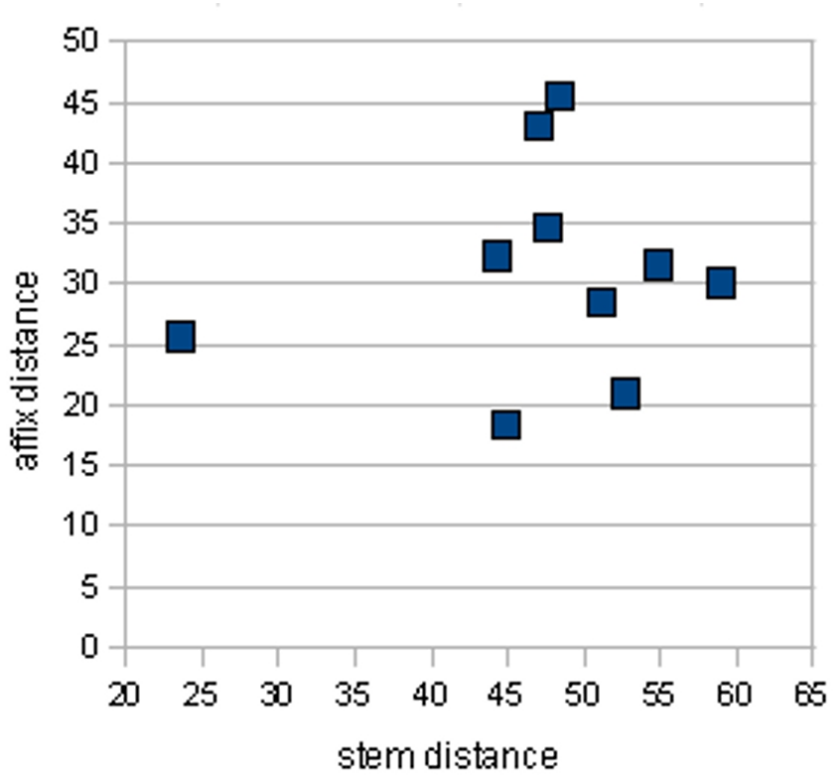
- Dendrogram obtained on the basis of stem distances (left) and affix distances (right):





First hypothesis

- Orthographic stem variation among languages does not correlate with orthographic variation in inflectional affixes.
- The hypothesis is true, since $r = 0.13$ and $p = 0.35$ (Mantel test)



First hypothesis

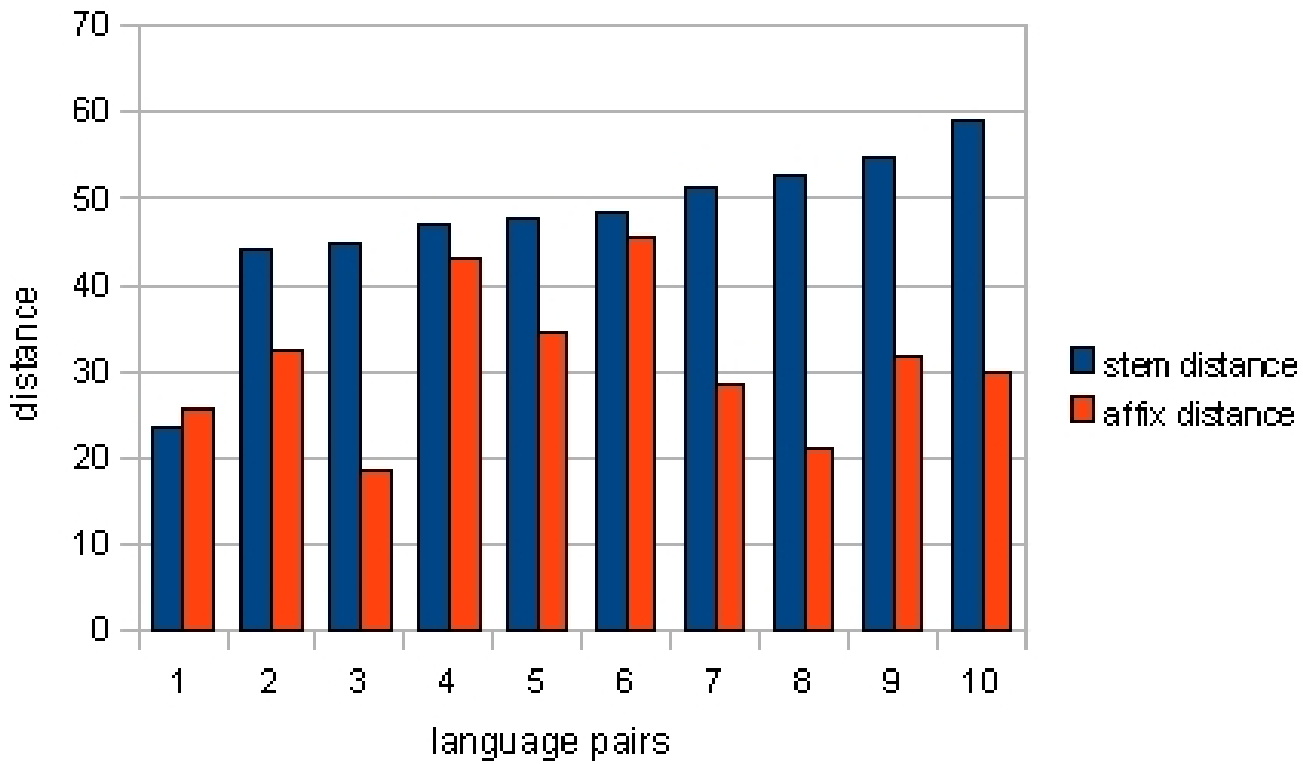
- Per language pair we correlate the stem pair distances with the affix pair distances. Results:

	Danish	Dutch	English	German	Swedish
Danish		0.16**	-0.25**	0.04	0.17**
Dutch			-0.20**	-0.14**	0.06
English				-0.05	-0.15**
German					0.06
Swedish					

- The correlations are low and/or not significant.

Second hypothesis

- Orthographic stem variation among languages is larger than orthographic variation in inflectional affixes.
- The hypothesis is true. Using a paired-samples t test we found: $t = 4.302$, $p = 0.001$ (one-sided), df=9.



Second hypothesis

- Per language pair it is tested whether the stem pair distances are higher than the affix pair distances with a paired-samples t-test. We obtained the following p-values:

	Danish	Dutch	English	German	Swedish
Danish		< 0.001	< 0.001	< 0.001	$= 0.358$
Dutch			< 0.001	< 0.001	$= 0.307$
English				< 0.001	< 0.001
German					$= 0.452$
Swedish					

Second hypothesis

- Affix pair distances are smaller than stem pair distances, except for most pairs in which Swedish is involved.
- Swedish is exceptional since the article is expressed as a suffix.

Conclusion

- When including orthography as a predictor in a model for language intelligibility, stem variation and affix variation should be distinguished.

Final remarks

More information can be found at:

`http://www.let.rug.nl/gooskens/project/`