Four approaches to automatic transcription of German names. A comparison.

Andreas Mengel, Technische Universitaet Berlin

At the Institut für Fernmeldetechnik Berlin four different methods for the purpose of automatic transcription of German names have been set up. All of these methods lack in reliability and performance. This paper describes an evaluation of the behavior of all of the four methods. A set of 1000 names was passed through all of them and checked for mistakes. Mistakes were counted in total and per category of mistake. Additionally, a common behavior of different methods is evaluated and used for the prediction of quality levels.

The transcription of names is a battle against an inconsistant and unsystematic codification of sound structures.

In Berlin, we have four systems available for the automatic transcription of names. These are: a rule-based system, a morphological system, a data-based system (Andersen 1994) and a neural net system (Rosenke 1994).

1 Data

For the evaluation of the performance of the system 1000 names of each of the categories christian names, surnames, street names and town names were selected randomly out of the data already transcribed. Table 1 gives some insight into the structure of the data.

name category	letters per entry	sounds per entry	sounds per letter
christian name	7,27	7,10	0,96
surname	8,94	7,76	0,87
town name	9,98	9,12	0,91
street name	12,17	10,99	0,90

Table 1: Structural statistics on names.

At first glance it seems as if the ratio sounds per letters increases (approximates zero) by the length. This is true for name categories that are more or less native. This rule does not hold for christian names which are more complex.

2 Procedure

The names selected were transcribed by each of the systems. The training set for the neural net and the data-based system consisted of 10810 christian names, 51478 surnames, 73622 street names and 25941 town names.

The transcriptions produced by the systems were automatically checked against hand prepared transcriptions. Errors were counted as mistranscribed entries as such and mistranscribed entries due to stress assignment errors, syllabification errors and sound errors. The last column in table 2 sums up the errors found. Sums that are higher than the number of mistranscribed names indicate that there are entries with more than one type of error. The columns of the morphological row in table 2 are empty as the morphological system either recognizes the

morphological structure of an entry or fails. i.e does not produce a transcription at all. Those names transcribed by the morphological system are nearly 100% errorfree.

2 Errors

Although there is no system that produces acceptable results - something that cannot be compared to publications of the performance of other systems as the transcription standards for ONOMASTICA are very ambitious including syllable boundaries and three levels of accentuation - the performance of the systems depends on the type of names. The morphological system has a poor perfomance on christian names that are not as rich in morphology as surnames and streetnames are.
The type or errors can also give information on the severeness of errors. Thus, sound errors are more likely to produce communication problems than stress errors will. From this recommadations can be drawn which procedure to take in order to reduce problems when using transcriptions that are likely to be wrong. A more detailed analysis can be found in Mengel & Rosenke (1994).

	entries	accents	boundaries	sounds	units
christian names					
rule based	49.9	1.2	4.9	44.7	50.8
data based	46.5	25.2	15.2	8.2	48.6
morphologic	88.1	-	-	-	-
neural net	71.2	4.9	12.4	55.8	73.1
surnames					
rule based	23.1	0.6	6.3	16.8	23.7
data based	27.2	12.3	7.2	14.3	33.8
morphologic	36.9	-	-	-	-
neural net	26.9	5.1	5.9	17.5	28.5
street names					
rule based	40.6	0.8	5.1	35.5	41.4
data based	49.4	33.9	25.7	14.8	74.4
morphologic	12.2	-	-	-	-
neural net	22.0	4	6.2	13.4	23.6
town names					
rule based	53.9	0.8	17.1	36.4	54.3
data based	43.2	24.9	20.5	11.1	56.5
morphologic	57.4	-	-	-	-
neural net	50.4	8.6	19.6	25.2	53.4

Table 2: Transcription errors [%]

3 Correct entries

If the task is to produce a lexicon, errors are not usefull it is of little help to know how many percent of the entries are correct or wrong but if we can identify which entries are correct

respectively wrong. To find out this is easy if a reference data base is at hand but in most cases it is not. Thus a heuristical approach was chosen: Names transcribed alike by two up to four systems were identified, correct and wrong cotranscriptions were counted (Table 3).

	system				combinations [%]		
name category	**rule**	**data**	**morph**	**neuro**	**total**	**correct**	**wrong**
christian names	●	●	●		32.7	32.6	0.1
	●		●		9.4	8.3	1.1
	●			●	25.6	22.7	2.9
		●	●		9.0	8.9	0.1
		●		●	22.7	22.7	-
			●	●	6.2	5.3	0.9
	●	●	●		6.1	6.1	-
	●	●		●	17.5	17.5	-
	●		●	●	4.2	3.7	0.5
		●	●	●	4.3	4.3	-
	●	●	●	●	3.0	3.0	-
surnames	●	●			56.4	56.4	-
	●		●		49.8	49.5	0.3
	●			●	59.9	58.2	1.7
		●	●		45.6	45.6	-
		●		●	56.5	56.3	0.2
			●	●	50.9	50.4	0.5
	●	●	●		35.8	35.8	-
	●	●		●	45.3	45.3	-
	●		●	●	39.7	39.6	0.1
		●	●	●	38.3	38.3	-
	●	●	●	●	30.3	30.3	-
street names	●	●			31.1	31.1	-
	●		●		54.9	54.5	0.4
	●			●	50.2	48.6	1.6
		●	●		46.4	46.4	-
		●		●	41.6	41.5	0.1
			●	●	71.7	71.4	0.3
	●	●	●		29.3	29.3	-
	●	●		●	25.6	25.6	-
	●		●	●	45.8	45.5	0.3
		●	●	●	38.6	38.6	-

	●	●	●	●	24.4	24.4	-
town names	●	●			33.5	33.4	0.1
	●		●		31.5	26.2	5.3
	●			●	34.2	30.5	3.7
		●	●		31.0	30.8	0.2
		●		●	39.1	38.7	0.4
			●	●	32.5	28.8	3.7
	●	●	●		20.2	20.2	-
	●	●		●	24.3	24.3	-
	●		●	●	19.7	18.4	1.3
		●	●	●	23.0	22.8	0.2
	●	●	●	●	15.4	15.4	-

Table 3: Cotranscriptions of names.

4 Conclusion

By the means of a database like table 3 one can predict the accuracy of transcriptions automatically produced. One can also gain information on the similarity of systems by the means of common errors they produce. Different systems that are to transcribe the names show different competence. Thus, we can claim that it is possible to produce error-free transcriptions without human control. This is what is needed dealing with 1 000 000 names that need to be transcribed automatically.

5 References

Andersen, Ove & Dalsgaard, Paul (1994): A Self Learning Approach to Transcription of Danish Proper Names. ICSLP Yokohama, 1627-1630.

Mengel, Andreas & Rosenke, Katrin (1994): Vergleich von Transkriptionsansätzen für deutsche Namen. Elektronische Sprachverarbeitung 1994, 453-459.

Rosenke, Katrin (1994): Einsatz von neuronalen Netzen zur Transkription von orthographischem Text in Lautschrift. Elektronische Sprachverarbeitung 1994, 460-467.