A COMPARISON OF GERMAN NAMES AND GERMAN WORDS

A. Mengel
Institute of Communications Research, Technische Universität, Berlin

ABSTRACT

German names are more difficult to read and write than German words. This paper presents evidence for this fact that allows for explanations of this behaviour by investigating the frequency, orthography, phonetic structure and interplay between written and spoken words. The consequences of the findings for future automated processing of names are recommended.

INTRODUCTION

It is commonly accepted that the relationship between the orthographic and phonetic structure of German names is more difficult to handle than that of German words. This has negative effects on speech synthesis and speech recognition systems on the one hand and on the use of German orthography and pronunciation by speakers on the other.

DATA

To substantiate and localise the reasons of the above experience, selected properties of four categories of proper names (christian names, surnames, street names, town names) and non-inflected nouns were compared. Non-inflected nouns were chosen, as they are that group of words which resemble names most in morphological, syntactical and semantical behaviour: They are not inflected, they can act as subjects and objects and they denote entities. Names are rarely inflected, thus, only non-inflected nouns were chosen. The names and their transcriptions were taken from the German part of a CD-ROM produced by LRE-Project of the European Community called ONOMASTICA [1]. The CD-ROM contains approximately 2,000,000 German proper names. Only data which had been checked by humans were chosen. The non-inflected nouns and their transcriptions were supplied by the CELEX lexical database [2]. Entries without frequency information were not taken into account.

The transcriptions of the data include information on the sounds, syllable boundaries, primary and secondary stress. The data from the CELEX were adapted to the transcription standard used in the ONOMASTICA data. Consonant and vowel clusters surrounding syllable boundaries were standardised across all kinds of entries.

Table 1 shows the categories chosen, the number of the entries taken, their cumulative frequency of occurrence, and the coverage of the selected data.

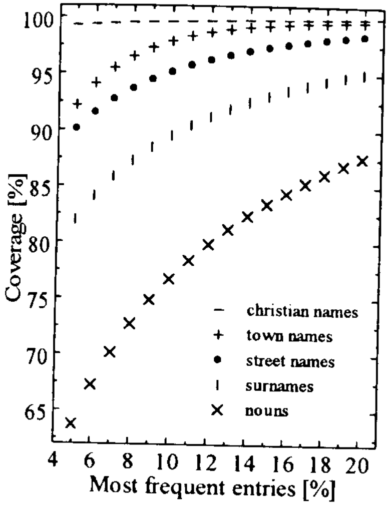
Table 1. Number, frequency and coverage of the selected data.

Cat.	n	Frequency	Cov. [%]
chri	10,778	33,040,984	97.62
sum	51,473	22,423,645	67.07
stre	73,605	34,182,103	63.82
town	25,892	40,570,131	99.34
noun	18,713	763,850	100.00

FREQUENCY OF OCCURRENCE

The number of different entries in the ONOMASTICA-corpus differs from category to category; also the number of items that are needed to reach a certain amount of coverage changes widely among types of entries. This can be seen best in figure 1. It shows how many of the most frequent entries of a name category are needed (x-axis) to cover a certain amount of entry frequency (y-axis). With 16% of the most frequent nouns, 85% of all occurrences of nouns are covered. The deviation of the graphs from an assumed straight line can be interpreted as a measure for the unevenness of distribution of entries in a given kind of name or the nouns and has the following implication: It is difficult to predict the occurrence of an entry in surnames or street names, but it will be easier in christian and town names because the likelihood for some entries is very high and is low for others. Thus, it seems that it is even more difficult to predict a certain noun than to predict a given surname. Yet, on the whole it will always be more difficult to predict a name than to predict a given word or noun: The usage of each individual noun in a given sentence is more restricted by syntactic and semantic constraints whereas there are less situations in which a name is predictable by the context.

Figure 1. Percentage of most frequent entries and coverage of entries.



LETTER-TO-SOUND RULES

It is difficult to measure the difference of correspondences between orthography and pronunciation across different name and word categories. Even more difficult is it to measure the extent of such differences. With rule-based systems at hand one can only argue on the basis of individual string-categories or special letter-sequences and their corresponding pronunciation. Thus, only non-rule-based approaches can function as an instrument for this sort of investigation. Two self-learning methods were applied to make a measurement possible. The first is a neural-net-based system (BACK), the net type of which is a multi-layer perceptron. It is trained by backpropagation [3]. The second system is a self-learning system [4] based on a statistical approach (SELEGRAPH). Both systems were independently trained with the five different data sets. The resulting nets (BACK) and databases (SELEGRAPH) should then represent five different functions between orthography and transcription of the five categories.

To investigate the common assumption that the transcription of words and different categories of names need separate sets of letter-to-sound rules, a set of identical character sequences must be transcribed by the five different versions of the two self-learning algorithms.

Six-thousand strings with lengths of four to six characters were identified. They can be found as substrings in entries of all of the five data types. Also, entries of these lengths can be found in each of the data types.

These 6,000 strings were transcribed by the 5 different versions of each system. Markers for suprasegmental information were deleted from the transcriptions. Then, each transcription produced by one version of the net/database (e.g. the BACK-system version trained for christian names) was compared to those produced by versions for other categories of names or nouns (i.e. the BACK-system versions for surnames, street-names etc.) .

Table 2 shows the percentage of the character sequences transcribed differently by a pair of differently trained systems for the neural net (BACK) and the statistical approach (SELEGRAPH). Figures of deviations are ordered in descending order.

Table 2: Percentage of differently transcribed strings.

BACK		SELEGRAPH	
diff. [%]	pairs	diff. [%]	pairs
60.55	chri-noun	40.23	chri-noun
57.45	chri-stre	34.87	stre-noun
50.78	chri-sum	34.78	sum-noun
50.52	chri-town	32.08	town-noun
44.97	stre-noun	31.60	chri-town
44.05	town-noun	31.11	chri-stre
41.58	sum-noun	28.85	chri-sum
37.53	sum-stre	16.43	sum-town
36.65	stre-town	16.15	stre-town
28.95	sum-town	14.65	sum-stre

Supposed, the percentage of differently transcribed entries is interpreted as measure for the difference between the LTS-correspondence of two separate entry types. It is then obvious that the difference of christian names and nouns is biggest (first row). However, the difference between surnames, street names and town names is smallest (last three rows). From the results, it is indeterminable whether the differences between christain names and the other name categories, or the difference between the nouns and the other names is bigger. Hence, christian names and nouns seem to mark the edges of the range of correspondences between German orthography and pronunciation.

MINIMAL PAIRS

The concept of minimal pairs is used to evaluate the phonetic similarity of a group of entries. Minimal pairs are pairs of words of equal number of sounds that differ from each other in one sound only. In this investigation, diphthongs were treated as long vowels, affricates as two sounds, and the glottal stop as a consonant. Table 3 shows the percentage of entries that have minimal-pair partners.

Table 3. Percentage of entries that have minimal-pair partners.

Category	Entries with minimal-pair partners [%]
chri	59.69
sum	70.98
stre	46.32
town	44.04
noun	19.58

All of the names have more minimal - pair partners than the nouns have. Thus, in a speech recognition system the recognition performance for nouns would be better than it would be for any name. It would be worst for surnames and christian names.

HOMONYMY

In order to have another look on the orthography of names and nouns, the number of different orthographic strings in the corpus under investigation is measured against the number of different phonetic strings. The ratio of different phonetic strings and different orthographic strings (p/o) is calculated.

Figure 4. Ratio of different phonetic and orthographic strings.

Category	p/o [%]
chri	87.86
sum	85.83
stre	92.19
town	97.61
noun	99.00

These results rather provide information on the relation of orthographic and phonetic structures of the entry types while minimal pairs only express something about the phonetic aspects of entries: The more homophones there are, i.e. orthographically different entries with the same pronunciation, the more difficult will it be to determine the correct orthography of a transcription or pronunciation. Hence, it is more difficult to find the correct entry. Again, this has severe impact on speech recognition.

CONCLUSION

Four aspects of differences between four types of German names and non-inflected nouns have been addressed: the frequency of occurrence, LTS-correspondences, the number of minimal pairs and homonymy. All of them show that especially surnames deserve particular attention and require more effort for processing, be it human or automatic. Thus, for the implementation of future applications that include the use of personal names, more refined methods must be developed to cope with the state-of-the-art performance achieved for words.

REFERENCES

[1] ONOMASTICA (1995): *Transcription database for proper names of 11 European languages,* Edinburgh: CCIR, University of Edinburgh.

[2] Baayen, R.H.; Piepenbrock, R. & van Rijn, H. (1993), *The CELEX Lexical Database (CD-ROM),* University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

[3] Rosenke, K. (1994), 'Einsatz von neuronalen Netzen zur Transkription von orthographischem Text in Lautschrift', *Konferenz 'Elektronische Sprachsignalverarbeitung'*, Technische Universität Berlin, Institut für Fernmeldetechnik, pp. 460-467.

[4] Andersen, O. & Dalsgaard, P. (1994) 'A Self Learning Approach to Transcription of Danish Proper Names', *Proceedings of the International Conference on Spoken Language Processing,* Yokohama, pp 1627-1630.