A Quantitative and Qualitative Analysis of Nordic Surnames

Eirini Florou
Institute of Informatics
and Telecommunications
NCSR 'Demokritos'
Athens, Greece
eirini.florou@gmail.com

Stasinos Konstantopoulos
Institute of Informatics
and Telecommunications
NCSR 'Demokritos'
Athens, Greece
konstant@iit.demokritos.gr

Abstract

Analysing Nordic persons' names with respect to language identification is a very hard task, as the chosen group of languages is closely related, but provides interesting insights into the structure of names; it is also a task that has many applications in information extraction, speech synthesis, and automatic transliteration. In this paper we present and discuss results obtained by statistical language modelling as well as by a handcrafted definite clause grammar.

1 Introduction

Language identification is the task of predicting the language that a text or utterance is written or spoken in. Language identification is typically approached as a statistical text categorization task, where features are extracted by analysing different linguistic levels, from the acoustic and prosodic to the phonotactic or graphotactic.

In this paper we concentrate on identifying the language of written text and, in fact, the language of a *single person's name*, in isolation or in a document written in a different language. This is a much harder task than predicting the language of texts, even short ones, but also one that is interesting from the theoretical as well as the practical perspective: from the *onomastics* perspective, it leads to interesting insights into the internal structure of names; from the *computational linguistics* perspective, it helps us explore the limits of various modelling methods on a task that is both hard and well-understood; and, finally, this task has interesting *natural language processing* applications in information extraction, speech synthesis, and automatic transliteration.

We further focus our task to the surnames of people from Denmark, Norway, Sweden, Finland, and Germany.[1] Surnames from Denmark, Norway, and Sweden form a cluster of particularly hard to separate surnames, while adding those from Finland and Germany leads to interesting conclusions as will be discussed later.

Our investigation builds on previous work on text categorization and, in particular, predicting the linguistic background of the bearer of a name (Section 2). To this end, we first apply statistical modelling (Section 3), the results of which guide us to propose a hypothesis (Section 4), which we validate by implementing as a definite clause grammar (Section 5). The paper closes drawing conclusions and outlining future work (Section 6).

2 Background

Guessing the language of a document falls under the larger area of *text categorization*, which aims at classifying a document as belonging to one (or more) out of certain, predefined categories or subject codes. Document language is one of the possible dimensions of categorization, interesting for various document organization, data mining, and information extraction tasks.

In their seminal paper, Cavnar and Trenkle (1994) report experiments on language categorization using a simple n-gram frequency algorithm. The language models consist of frequency counts of n-grams (up to 5-grams) for various languages. To classify a document, the frequency counts of n-grams in the document are calculated and their distribution compared against the distribution of n-grams in the language models. The model with the smallest distance from the distribution of the document, is assumed to be the language of the document.

[1] In the remainder of this paper we, somewhat arbitrarily and for lack of a perfectly-fitting term, collectively refer to these five classes as *Nordic*. We, furthermore, occasionally abbreviate them, in tables and elsewhere, as follows: DA (Denmark), NO (Norway), SV (Sweden), FI (Finland), and DE (Germany).

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 74–81

This algorithm was tested on Usenet postings from the soc.culture newsgroup hierarchy. An eight-language corpus was generated semi-automatically: a first pass operated under the assumption that the postings are in the language of the country or region under discussion in each newsgroup, and at a second pass discrepancies between the newsgroup's default language and the system's prediction were manually resolved.

With the 400 most frequent n-grams retained in the models, and postings of at least 300 characters of length, the system classified the test set almost perfectly, achieving an accuracy of 99.8%. The authors also report an accuracy of 99.3% for postings that are under 300 characters, without providing any further details of how accuracy drops with shorter test documents.

Cavnar and Trenkle's algorithm has seen various implementations and applications, the most notable probably being the TEXTCAT[2] implementation used in the SPAMASSASSIN[3] spam filter.

Although very accurate even for texts as small as two or three hundred characters, Cavnar and Trenkle's experiments did not test how well one can identify the language of a single word. Efforts in this, much harder, task originate in speech synthesis (Spiegel, 1985; Vitale, 1991; Font Llitjós and Black, 2001), with language identification used to select different pronunciation models for foreign names, depending on each name's origin.

Font Llitjós and Black (2001), in particular, note that language identification of isolated names is a difficult task, as they tried to manually tag 516 names and found that they could confidently tag only 43% of the data. For their speech synthesis experiment they used a simplification of the Cavnar and Trenkle algorithm which only counted 3-grams. They trained language models on general text (ranging from 255 thousand to 11 million words), and provided the classification results as features for the grapheme-to-phoneme models. Unfortunately they do not report results for the language identification part of their experiments.

Another field of application of the same general methodology is automatic transliteration of named-entities for the purposes of *information extraction* (Virga and Khudanpur, 2003) or *machine translation* (Huang, 2005). In Huang's experiment languages were grouped together in clusters, guided by the effect each clustering had on the accuracy of the overall transliteration. The resulting clusters roughly corresponded to familiar language groupings (Chinese, Romance, English-and-Dutch, Nordic). Employing language identification models is reported to improve the accuracy of the overall task, but no results are provided for the language identification sub-task per se. Virga and Khudanpur (2003) report improved accuracy in recovering the original orthography of English-language named-entities in Chinese text by using a tri-gram model to first decide whether a Chinese string is an English-language name or not or selecting different transliteration model depending on his.

More recently, Konstantopoulos (2007) presented a corpus of European names compiled by harvesting information from the web. The corpus matches about 15 thousand names against their nationality and was used to compare the accuracy of n-gram modelling on language identification of a single name against language identification of common words of the same size, finding the former to offer themselves to significantly more accurate prediction. Konstantopoulos (2010) followed up with applying a series of statistical tests looking for the discriminative features in names that offer themselves to more accurate prediction, but without reaching any definite conclusions.

3 Statistical Modelling

As a first, exploratory, step we applied the TEXTCAT implementation of Cavnar and Trenkle's 5-gramm modelling method and tested the language models of the training set itself to derive confusion matrices. For this, we used the relevant parts of the corpus created by Konstantopoulos (2007). This corpus was created by harvesting the Transfermarkt website,[4] featuring various information about football players, including—most crucially for our purposes—their nationality. We extended this corpus with names harvested from lists of members of parliament. Because of the strictness of football associations' naturalization rules and the increased inertia in politics at the national level, these names delineate tightly, although not perfectly, the linguistic background of their bearers. In total, we were able to compile a corpus of 5568 names, distributed

[2] http://www.let.rug.nl/~vannoord/TextCat/

[3] http://spamassassin.apache.org/

[4] See http://www.transfermarkt.de/

Table 1: Confusion matrix of surnames, given as the fraction of predictions that a string is in each class (rows) for each of the actual classes (columns), so that the numbers along the diagonal (in bold) represent the *recall* achieved in each class. The absolute size of each class in the dataset is given in the bottom row of the table.

		Actual				
		DE	DA	NO	SV	FI
Predicted	DE	**0.73**	0.24	0.13	0.16	0.07
	DA	0.10	**0.54**	0.16	0.06	0.04
	NO	0.10	0.17	**0.63**	0.16	0.10
	SV	0.00	0.01	0.01	**0.64**	0.01
	FI	0.07	0.04	0.06	0.04	**0.78**
	Size	2608	678	987	629	666

among the five languages under consideration as shown on the last row of Table 1.

With respect to the choice of languages, we have chosen Swedish, Norwegian, and Danish because the form a cluster of closely related languages with similar orthographic conventions, making language identification a challenging task. German and Finnish were included in order to study interactions at the periphery of the three core target languages.

It should be noted that, since TEXTCAT language models comprise absolute frequency counts, we reduplicated names from the less populous name lists in order to have a balanced distribution of instances among classes and avoid having n-grams that are even moderately (in relative terms) frequent in German names dominate relatively frequent n-grams from other languages. We should also note that we have reduced all letters with diacritics to their plain Latin base letter. This was done for two reasons: in order to normalize orthography into a form that minimizes 'easy guesses' of the ø/ö kind, and in order to have an approach that fits named-entity recognition tasks in foreign-language contexts, where diacritics are often simply omitted, rather than transcribed into digraphs. That is, in such applications *å* typically becomes *a* rather than *aa*.

4 Discriminative Features

As Table 1 shows, this is a very hard task and the direct application of language modelling does not get us very hard, even when testing over the training set. But, as noted before, this was only an exploratory stage, where the observation of the results and the frequencies recorded in the models helped us identify the features that best discriminate surnames as well as those that cause the most confusion. From this basis, we looked for further morphological and semantic features that, we postulate, improve discernability; we used these extra features to formulate a series of hypotheses about the structure of Nordic surnames, presented in this section and tested in Section 5.

4.1 Patronymic surnames

Arguably the most characteristic Nordic surnames are patronyms ending in *–sen* (DA, NO) or *–son* (SV) suffixed to a first name. As *–sen* is shared between Danish and Norwegian, accounting for 26% of Danish and 31% of Norwegian surnames, and since there is very little (if any) grounds for correctly classifying the first names, this explains most of the confusion between these two languages' surnames seen on Table 1.

Swedish patronyms, on the other hand, can be easily spotted as they follow a different pattern where the first name in genitive (marked by *–s*) is followed by *–son*. This leaves no margin for ambiguity within the scope of our experiments, as 'son' is spelt *Sohn* in German and is rarely found in surnames anyway. Even beyond the scope of this paper, Scottish surnames that often exhibit this pattern should be relatively easy to separate by first name.

4.2 Toponymic surnames

Ending in *–er* is a common feature of all Nordic surnames, in fact, words in general, and accounts for 17.4% of German, 6.6% of Danish, 4.3% of Swedish, 2.7% of Norwegian, and 1.2% of Finnish surnames. Although of low discriminative power by itself, the *–er* suffix is important in our observations when co-occurring with other features.

In our data, surname derivations are mostly applied to monomorphemic roots with the exception of German where surnames are often derived from roots that are themselves derivatives. More specifically, 362 out of the 2608 German surnames in our corpus exhibit this property, typically derived from placenames; *Pfeifenberger*, *Rasswalder*, and *Amerhauser* are characteristic examples.

A more marked observation, also typical of German surnames, is that they can be derived

Table 2: Some examples of surnames including the n-grams <ander> and <land>.

Swedish	Danish/Norwegian	German
*Ander*sson	*Ander*sen	*Ander*sohn
Se*lander*		L*ander*l
	Klit*land*/Hel*land*	Wei*land*

from surnames, as exemplified by *Husterer* and *Eibinger*.

A characteristic failure of our language models is with toponymic surnames where the placename ends in *–land*. In Swedish it is common to use *–er* to derive such toponymic surnames (example shown in second row of Table 2), adding considerably to the frequencies of the <ander> and <land> n-grams in the Swedish language model. As 5-gram models are unable to capture the distinction between <lander> and <ander>, this feature causes several misclassifications, as <ander> appears frequently in Danish, Norwegian, and German surnames. Some characteristic examples are shown on Table 2.

The problem is further aggravated by the frequent appearance of <land> across all four languages. For example, <land\$>[5] is less characteristic of Swedish (i.e., more common across all languages) than <lander\$>, which should make <land\$> a weaker indicator than <lander\$>. Because, however, <lander\$> subsumes <land\$>, the latter's weight in the Swedish language model is boosted causing misclassifications.

There are a lot of cases of Danish surnames that the *–er* occurs in the morpheme *–ager*. This suffix can be added to whatever monosyllabic lexeme of the same language which could be Danish first or last name.

Also, Nordic surnames, except for Finnish, are formed by adding the *–er* to each language's corresponding toponyms in order to denote the person who is descended from the certain local. A series of examples indicates this fact.

4.3 German and Danish surnames

Occupational surnames are one of the major factors of the misclassification of many Danish surnames as German. As one can observe on Table 1, the confusion is largely asymmetric, the reason being the existence of many German-language occupational surnames in the Danish data, with the German orthography retained. One possible explanation is that these names were incorporated in Danish in the sixteenth and seventeenth centuries by the immigration of travelling guilds from Germany, who had already adopted occupational surnames.

Although the confusion can be somewhat alleviated by the fact that compounds such as *Rothbauer* and *Sommermayer* are only found in the German data, it is impossible to correctly classify surnames such as *Möller*, *Weber*, *Meyer*, and *Schmidt* found in the Danish data.

Another factor of confusion between German and Danish surnames, but also one that helps in the hard task of separating Danish from Norwegian, is the suffix *–ing* is almost exclusively found in German (2%) and Danish (2.5%). A differentiating factor is that in German *–ing* is sometimes part of a longer, uniquely German, morpheme such as *–ling* (e.g., *Emmerling*), and reduplication that might sometimes manifest this 4-gram can be easily spotted (as in, e.g., Danish *Balling*). Furthermore, a large class of German surnames in *–ing* are occupational surnames (e.g., *Möllering*) where Danish ones are placenames in Denmark (e.g., *Gjesing*, *Grønning*).

4.4 Compounding

Many Nordic surnames are formed by compounding a nominal modifier with natural or man-made features, such as *berg* 'mountain' (DE, SV) or 'iceberg' (DA) or *gaard* 'farm'. In the case of *berg*:

<nominal> + *berg*

where <nominal> is of the same language as the surname and in the case of Norwegian and German surnames can also be an adjective whereas in Swedish and Danish it is always a noun, possibly marked for plural or genitive (especially in Danish). Some indicative examples are given on Table 3. This is a fairly common pattern, matching 6% of the Swedish data and 2% of the Danish, Norwegian and German data. It should be noted that *Berg* also appears as the modifier in compounds such as *Bergqvist* (SV) or *Bergheim* (DE).

Another common pattern in Denmark and Norway is compounds with with *gaard*, accounting for 7% of Danish surnames, 4.5% of Norwegian sur-

[5] We use <\$> to signify word boundaries, consuming one of the 5 characters of each entry in the language model.

Table 3: Examples of surnames in *–berg*

Norwegian	German	Swedish	Danish
Ny-*berg*	Stolz-en-*berg*	Ceder-*berg*	Co-*berg*
Skjon-s-*berg*	Grun-*berg*	For-s-*berg*	Falken-*berg*

Table 4: Surnames in *–gaard*

Danish	Norwegian
Abil*gaard*	Ost*gaard*
Bis*gaard*	Ny*gaard*
Son*gaard*	Kort*gaard*

names, and many misclassifications. As is the case with *berg*, Norwegian surnames can be spotted by adjective modifiers, whereas Danish only allows nouns, typically toponyms.

Moreover, surnames may be formed by compounding with *man(n)*, where German/Danish use *mann* and Swedish *man*. The percentages of Nordic surnames with the certain suffix are respectively: DA 2%, DE 4%, SV 3%. Furthermore, German (but not Danish) *mann* compounds are ofter formed with a derived modifier (e.g., *–er-mann*, *–el-mann*), helping us identify surnames like *Kellermann* as German. Simpler surnames such as Danish *Hermann* can be easily misclassified.

Another common pattern, setting Swedish surnames apart, is that Swedish compound surnames often use roots related to natural features, for instance *Lind* 'lime tree', *Ceder* 'cedar'.

4.5 Some observations on *–er* surnames

The n-gram analysis has shown that there is a series of n-grams which contribute to Nordic surnames ambiguation. However, there are other features which can help to cause disambiguation.

Typical case is the bigram *–er* which is a common feature of all Nordic surnames. But only German surnames can be derived from other already existed surnames while all Nordic surnames are formed by adding *–er* to corresponding toponyms. Moreover, the occupational surnames are very common in German and Danish language. However, there are German surnames which can be formed by adding an occupational name to another typical German first or last name. Furthermore, there are typical per language morphemes which include the certain suffix and as a consequence the morpheme can bring better recognition results in contrast with the single suffix.

Apart from few exceptions, the majority of n-grams which cause Nordic surnames confusion have similarities with *–er* in their ability to identify the Nordic language in combination with other discriminative features.

5 Rule Modelling

Through statistical language modelling, we postulated a hypothesis about discriminative features of Nordic surnames that cannot be captured by 5-grams, either because the required context is too long or because they refer to morphological or semantic background knowledge.

As such dependencies and background can be more intuitively expressed as *definite clause grammars (DCG)*, we decided to use this framework to formalize and test out hypotheses. Besides long-distance dependencies and the ease of incorporation of external background knowledge, DCG also provide a declarative, intuitive representation that can more easily maintained and extended with new linguistic constructs.

This section presents the grammar and the results obtained over the data. Apart from the rules which are described below and which, arguably, capture general linguistic structures, our grammar also relies on a lexicon of entity names (first names, place names) and a lexicon of common words with part-of-speech annotations (adjectives, nouns, adverbs) as well as limited semantic annotations (e.g., words denoting geographic features). These resources were partially automatically created: regarding entity names we collected first names from the Konstantopoulos (2007) corpus and from a baby-names website[6] and toponyms from the Geonames database[7]. Other semantic classes have been manually compiled.

5.1 Some characteristic rules

As shown in the previous section, reference to part-of-speech and semantic class can correctly classify many difficult instances in our data. For instance, Norwegian and Danish *–gaard* compounds differ in that only Norwegian surnames use adjectives to modify *gaard*.

This is captured by the rules:

[6] http://www.babynames.com

[7] http://www.geonames.org

```
name(no)→lex(adj,_) gaard
name(da)→lex(plce(da),_) gaard
name(L) →lex(_,L) gaard
```

The second argument of `lex/2` encodes the prediction obtained by n-gram modelling and the `plce(Country)` lexical category signifies the placenames acquired from the Geonames database. With these rules, we can override the n-gram prediction when part-of-speech or semantics (Danish placename) offers itself for a more certain guess. Another hard case, *berg* compounds, is treated by generalizing the rules above into compounding rules such as:[8]

```
name(no)→lex(adj,_) lex(geofeat,no)
name(de)→lex(adj,_) lex(geofeat,de)
name(da)→lex(plce(da),_) lex(geofeat,da)
name(L) →lex(_,L) lex(geofeat,L)
```

where lexical category `geofeat` includes words such as *gaard* and *berg*.

Other rules override language modelling predictions in situations where n-grams, such as <ander>, are known to cause misclassifications in certain contexts. Referring to Section 4.2 above:

```
name(sv)→lex(_,_) land er
```

overrides any language modelling prediction made about the origin of the name. Surname-surname derivations can be recursively stipulated to be German irrespective of what the root surname is categorized as:

```
name(de)→name(_) er
```

5.2 Results

The n-gram modelling results (cf. Section 3), as well as the observation of the frequencies in the language models themselves, helped us identify discriminative and confusing features and drove our investigation into structure of surnames, culminating into a definite clause grammar that formalizes our observations.

This allowed a combination of n-grams results with morphological and semantic features that could be captured with n-gram modelling. Performance was evaluated in terms of *precision*, *recall* and the combined *F-measure*, as shown on Table 5. It should be noted that the test set is the training set itself, so that the predictive accuracy of the models is expected to be lower, but what is mostly relevant is the significant increase that the DCG achieves. The only single measurement that is higher in the n-gram modelling experiment is precision over Swedish names, which is clearly due to the overspecificity of the model, as demonstrated by the low recall and the fact that the better balance between precision and recall in the grammar achieves a much higher F-measure.

Table 5: Recall R, precision P and F-measure $F = 2RP/(R+P)$ obtained by the n-gram model and the definite clause grammar.

	n-gram model			DCG		
	R	P	F	R	P	F
DE	0.73	0.54	0.62	0.86	0.79	0.82
DA	0.54	0.57	0.55	0.72	0.58	0.64
NO	0.63	0.57	0.60	0.64	0.92	0.75
SV	0.50	0.94	0.65	0.63	0.84	0.72
FI	0.78	0.77	0.77	0.78	0.89	0.84

Besides these qualitative results, we have shown above examples of rules and given qualitative explanations of why they capture patterns that cannot be modelled with n-grams. Furthermore, we should also note that the DCG is a sounder generalization of the data as, notwithstanding the lexical resources it has access to, it is considerably shorter (just below 60 rules comprising fewer than 200 terms) than the 400 n-grams *per language* that the statistical language models retain. The lexical resources (geographic names and terms, part-of-speech annotations) are, naturally, very voluminous but one can arguably claim that they are not part of these experiments' foreground as their are independently justified semantic and morphological annotations. This brings forwards one of the most advertised advantages of DCGs and explicit representations in general, that is, their ability to exploit background knowledge.

Another decisive factor is the ability to express long-distance dependencies. Consider, for example, our DCG treatment of surnames in *–er*, a suffix that occurs frequently across all investigated languages and is not, by itself, a useful discriminator: the characteristically German surname-from-surname derivation, the identification of larger contexts such as Swedish *–land-er*, and other similar longer-than-five-gram features have dramatically improved the performance of these surnames.

[8] We have taken some liberties with non-essential details of rules for the purpose conciseness and clarity of the presentation. The interested reader can contact either author to obtain the full grammar.

6 Conclusions and Future work

In this paper we approached the very hard problem of identifying the linguistic background of Nordic surnames from two angles: as a statistical machine learning task and as a grammar engineering task. The main contributions are providing the resources and establishing a methodology for formulating, formalizing, and testing hypotheses about name structure; demonstrating a concrete case where the combination of machine learning and grammar engineering has proved beneficial; and, finally, improving performance on a very hard task that can been used on a variety of natural language processing applications.

Our methodology proposes using statistical language modelling to 'map' the domain, examining the n-grams in the language models to identify those that are good discriminators and, most critically, misleading n-grams that, although frequent, often lead to misclassifications. Building on these insights, a definite clause grammar is created which analyses surnames in order to classify them. This grammar has access to morpheme annotations derived from a variety of resources and analysis tools, including graphotactic (the n-gram modelling predictions), morphological (part-of-speech annotations), and semantic (the Geonames hierarchy).

We have measured the performance of the definite clause grammar to be significantly higher than that achieved by statistical language modelling alone, demonstrating that it can make good use of the additional morphological and semantic background that was available.

Besides the methodology itself, our contribution extends to discovering certain aspects of the structure of Nordic surnames that have, to the best of our knowledge, not been previously reported, such surnames being derived from surnames being a characteristically German trait. We have also augmented the Konstantopoulos (2007) names corpus with further entries.[9]

Naturally, there is considerable room for further work. Initially, it would be interesting to examine name structure at a purely phonotactic level where no guesses can be made from orthographic conventions. Besides the theoretical interest, this would also make our work relevant to information extraction from speech and spoken language machine translation tasks, where the orthography (even the simplified orthography we based our experiments on) is not recoverable before having recognized the language of the name.

It would also be interesting to compare grammars generated by a machine learning method, such as *inductive logic programming*, against our grammar, as well as surname recognizers against recognizers of other nominal classes, such as compounds. Such a survey would help us understand which of the structures we have discovered are characteristic of surnames and which are reflections of more general word-formation phenomena in these languages.

Acknowledgements

Stasinos Konstantopoulos wishes to acknowledge the support of the FP7-ICT project PRONTO.[10]

PRONTO develops methodologies for the analysis and interpretation of textual, audio, and video data, aiming at the extraction of operational knowledge supporting and improving resource management. In this context, the work described here is applied to the treatment of out-of-dictionary words in text and in audio transcriptions.

References

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR 94), Las Vegas, 11–13 April 1994*, pages 161–175.

Ariadna Font Llitjós and Alan W. Black. 2001. Knowledge of language origin improves pronunciation accuracy of proper names. In *Proc. of Eurospeech 2001, Aalborg, Denmark.*

Fei Huang. 2005. Cluster-specific named entity transliteration. In *Proceedings Conf. on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, British Columbia, Canada*, pages 435–442.

Stasinos Konstantopoulos. 2007. What's in a name? In *Proc. RANLP Workshop on Computational Phonology, Borovets, Bulgaria, September 2007.*

[9] The augmented corpus and various scripts for its manipulation are available at `http://www.iit.demokritos.gr/~konstant/dload/tmc.tgz`

[10] See `http://www.ict-pronto.org`

Stasinos Konstantopoulos. 2010. Learning language identification models: a comparative analysis of the distinctive features of names and common words. In *Proc. 7th Intl Conf. on Language Resources and Evaluation (LREC-2010), 19–21 May, Valletta, Malta*, pages 3431–6.

Murray F. Spiegel. 1985. Pronouncing surnames automatically. In *Proc. Conf. of the American Voice Input/Output Society*, pages 109–132.

Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of proper names in cross-lingual information retrieval. In *Proc. ACL Workshop on Multilingual and Mixed-language Named Entity Recognition*, pages 57–64.

Tony Vitale. 1991. An algorithm for high accuracy name pronunciation by parametric speech synthesizer. *Computational Linguistics*, 17(3):257–276.