Reduced *n*-gram models for English and Chinese corpora

Le Q Ha, P Hanna, D W Stewart and F J Smith
School of Electronics, Electrical Engineering and Computer Science,
Queen's University Belfast
Belfast BT7 1NN, Northern Ireland, United Kingdom
lequanha@lequanha.com

Abstract

Statistical language models should improve as the size of the *n*-grams increases from 3 to 5 or higher. However, the number of parameters and calculations, and the storage requirement increase very rapidly if we attempt to store all possible combinations of *n*-grams. To avoid these problems, the reduced *n*-grams' approach previously developed by O'Boyle (1993) can be applied. A reduced *n*-gram language model can store an entire corpus's phrase-history length within feasible storage limits. Another theoretical advantage of reduced *n*-grams is that they are closer to being semantically complete than traditional models, which include all *n*-grams. In our experiments, the reduced *n*-gram Zipf curves are first presented, and compared with previously obtained conventional *n*-grams for both English and Chinese. The reduced *n*-gram model is then applied to large English and Chinese corpora. For English, we can reduce the model sizes, compared to 7-gram traditional model sizes, with factors of 14.6 for a 40-million-word corpus and 11.0 for a 500-million-word corpus while obtaining 5.8% and 4.2% improvements in perplexities. For Chinese, we gain a 16.9% perplexity reductions and we reduce the model size by a factor larger than 11.2. This paper is a step towards the modeling of English and Chinese using semantically complete phrases in an *n*-gram model.

1 Introduction to the Reduced *N*-Gram Approach

Shortly after this laboratory first published a variable *n*-gram algorithm (Smith and O'Boyle, 1992), O'Boyle (1993) proposed a statistical method to improve language models based on the removal of overlapping phrases.

A distortion in the use of phrase frequencies had been observed in the small railway timetable Vodis Corpus when the bigram "RAIL ENQUIRIES" and its super-phrase "BRITISH RAIL ENQUIRIES" were examined. Both occur 73 times, which is a large number for such a small corpus. "ENQUIRIES" follows "RAIL" with a very high probability when it is preceded by "BRITISH." However, when "RAIL" is preceded by words other than "BRITISH," "ENQUIRIES" does not occur, but words like "TICKET" or "JOURNEY" may. Thus, the bigram "RAIL ENQUIRIES" gives a misleading probability that "RAIL" is followed by "ENQUIRIES" irrespective of what precedes it. At the time of their research, O'Boyle reduced the frequencies of "RAIL ENQUIRIES" by subtracting the frequency of the larger trigram, which gave a probability of zero for "ENQUIRIES" following "RAIL" if it was not preceded by "BRITISH." The phrase with a new reduced frequency is called a reduced phrase.

Therefore, a phrase can occur in a corpus as a reduced *n*-gram in some places and as part of a larger reduced *n*-gram in other places. In a reduced model, the occurrence of an *n*-gram is not counted when it is a part of a larger reduced *n*-gram. One algorithm to detect/identify/extract reduced *n*-grams from a corpus is the so-called reduced *n*-gram algorithm. In 1992, O'Boyle was able to use it to analyse the Brown corpus of American English (Francis and Kucera, 1964) (of one million word tokens, whose longest phrase-

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 309–315,
Sydney, July 2006. ©2006 Association for Computational Linguistics

length is 30), which was a considerable improvement at the time. The results were used in an *n*-gram language model by O'Boyle, but with poor results, due to lack of statistics from such a small corpus. We have developed and present here a modification of his method, and we discuss its usefulness for reducing *n*-gram perplexity.

2 Similar Approaches and Capability

Recent progress in variable *n*-gram language modeling has provided an efficient representation of *n*-gram models and made the training of higher order *n*-grams possible. Compared to variable *n*-grams, class-based language models are more often used to reduce the size of a language model, but this typically leads to recognition performance degradation. Classes can alternatively be used to smooth a language model or provide back-off estimates, which have led to small performance gains. For the LOB corpus, the varigram model obtained 11.3% higher perplexity in comparison with the word-trigram model (Niesler and Woodland, 1996.)

Kneser (1996) built up variable-context length language models based on the North American Business News (NAB-240 million words) and the German Verbmobil (300,000 words with a vocabulary of 5,000 types.) His results show that the variable-length model outperforms conventional models of the same size, and if a moderate loss in performance is acceptable, that the size of a language model can be reduced drastically by using his pruning algorithm. Kneser's results improve with longer contexts and a same number of parameters. For example, reducing the size of the standard NAB trigram model by a factor of 3 results in a loss of only 7% in perplexity and 3% in the word error rate. The improvement obtained by Kneser's method depended on the length of the fixed context and on the amount of available training data. In the case of the NAB corpus, the improvement was 10% in perplexity.

Siu and Ostendorf (2000) developed Kneser‘s basic ideas further and applied the variable 4-gram, thus improving the perplexity and word error rate results compared to a fixed trigram model. They obtained word error reductions of 0.1 and 0.5% (absolute) in development and evaluation test sets, respectively. However, the number of parameters was reduced by 60%. By using the variable 4-gram, they were able to model a longer history while reducing the size of the model by more than 50%, compared to a regular trigram model, and at the same time improve both the test-set perplexity and recognition performance. They also reduced the size of the model by an additional 8%.

Other related work are those of Seymore and Rosenfeld (1996); Hu, Turin and Brown (1997); Blasig (1999); and Goodman and Gao (2000.)

In order to obtain an overview of variable *n*-grams, Table 1 combines all of their results.

COMBINATION OF LANGUAGE MODEL TYPES								
Basic *n*-gram	**Variable *n*-grams**	**Category**	**Skipping distance**	**Classes**	**#params**	**Perplexity**	**Size**	**Source**
Trigram√					987k	474	1M	LOB
		Bigram√			-	603.2		
		Trigram√			-	544.1		
	√	√			-	534.1		
Trigram√					743k	81.5	2M	Switch board Corpus
	Trigram√				379k	78.1		
	Trigram√		√		363k	78.0		
	Trigram√		√	√	338k	77.7		
	4-gram√				580k	108		
	4-gram√		√		577k	108		
	4-gram√		√	√	536k	107		
	5-gram√				383k	77.5		
	5-gram√		√		381k	77.4		
	5-gram√		√	√	359k	77.2		

Table 1. Comparison of combinations of variable *n*-grams and other Language Models

3 Reduced *N*-Gram Algorithm

The main goal of this algorithm (Ha, 2005) is to produce three main files from the training text:

- The file that contains all the complete *n*-grams appearing at least *m* times is called the PHR file ($m \geq 2$.)
- The file that contains all the *n*-grams appearing as sub-phrases, following the removal of the first word from any other complete *n*-gram in the PHR file, is called the SUB file.
- The file that contains any overlapping *n*-grams that occur at least *m* times in the SUB file is called the LOS file.

The final list of reduced phrases is called the FIN file, where

$$FIN := PHR + LOS - SUB \quad (1)$$

Before O'Boyle's work, a student (Craig) in an unpublished project used a loop algorithm that was equivalent to FIN:=PHR–SUB. This yields negative frequencies for some resulting *n*-grams with overlapping, hence the need for the LOS file.

There are 2 additional files

- To create the PHR file, a SOR file is needed that contains all the complete *n*-grams regardless of *m* (the SOR file is the PHR file in the special case where $m = 1$.) To create the PHR file, words are removed from the right-hand side of each SOR phrase in the SOR file until the resultant phrase appears at least *m* times (if the phrase already occurs more than *m* times, no words will be removed.)
- To create the LOS file, O'Boyle applied a POS file: for any SUB phrase, if one word can be added back on the right-hand side (previously removed when the PHR file was created from the SOR file), then one POS phrase will exist as the added phrase. Thus, if any POS phrase appears at least *m* times, its original SUB phrase will be an overlapping *n*-gram in the LOS file.

The application scope of O'Boyle's reduced *n*-gram algorithm is limited to small corpora, such as the Brown corpus (American English) of 1 million words (1992), in which the longest phrase has 30 words. Now their algorithm, re-checked by us, still works for medium size and large corpora. In order to work well for very large corpora, it has been implemented by file distribution and sort processes.

Ha, Seymour, Hanna and Smith (2005) have investigated a reduced *n*-gram model for the Chinese TREC corpus of the Linguistic Data Consortium (LDC) (http://www.ldc.upenn.edu/), catalog no. LDC2000T52.

4 Reduced *N*-Grams and Zipf's Law

By re-applying O'Boyle and Smith's algorithm, we obtained reduced *n*-grams from two English large corpora and a Chinese large corpus.

The two English corpora used in our experiments are the full text of articles appearing in the Wall Street Journal (WSJ) (Paul and Baker, 1992) for 1987, 1988, 1989, with sizes approximately 19 million, 16 million and 6 million tokens respectively; and the North American News Text (NANT) corpus from the LDC, sizing 500 million tokens, including Los Angeles Times & Washington Post for May 1994-August 1997, New York Times News Syndicate for July 1994-December 1996, Reuters News Service (General & Financial) for April 1994-December 1996 and Wall Street Journal for July 1994-December 1996. Therefore, the WSJ parts from the two English corpora are not overlapping together.

The Mandarin News corpus from the LDC, catalog no. LDC95T13 was obtained from the People's Daily Newspaper from 1991 to 1996 (125 million syllables); from the Xinhua News Agency from 1994 to 1996 (25 million syllables); and from transcripts of China Radio International broadcast from 1994 to 1996 (100 million syllables), altogether over 250 million syllables. The number of syllable types (i.e. unigrams) in the Mandarin News corpus is 6,800. Ha, Sicilia-Garcia, Ming and Smith (2003) produced a compound word version of the Mandarin News corpus with 50,000 types; this version was employed in our study for reduced *n*-grams.

We next present the Zipf curves (Zipf, 1949) for the English and Chinese reduced *n*-grams.

4.1 Wall Street Journal

The WSJ reduced *n*-grams can be created by the original O'Boyle-Smith algorithm implemented on a Pentium II 586 of 512MByte RAM for over 40 hours, the disk storage requirement being only 5GBytes.

The conventional 10-highest frequency WSJ words have been published by Ha et al. (2002) and the most common WSJ reduced unigrams, bigrams and trigrams are shown in Table 2. It illustrates that the most common reduced word is not THE; even OF is not in the top ten. These words are now mainly part of longer *n*-grams with large *n*.

The Zipf curves are plotted for reduced unigrams and *n*-grams in Figure 1 showing all the curves have slopes within [-0.6, -0.5]. The WSJ reduced bigram, trigram, 4-gram and 5-gram curves become almost parallel and straight, with a small observed noise between the reduced 4-gram and 5-gram curves when they cut each other at the beginning. Note that information theory tells us that an ideal information channel would be made of symbols with the same probability. So having a slope of –0.5 is closer than –1 to this ideal.

Rank	Unigrams		Bigrams		Trigrams	
	Freq	Token	Freq	Token	Freq	Token
1	4,273	Mr.	2,268	he said	1,231	terms weren’t disclosed
2	2,469	but	2,052	he says	709	the company said
3	2,422	and	1,945	but the	664	as previously reported
4	2,144	the	1,503	but Mr.	538	he said the
5	1,918	says	1,332	and the	524	a spokesman for
6	1,660	or	950	says Mr.	523	the spokesman said
7	1,249	said	856	in addition	488	as a result
8	1,101	however	855	and Mr.	484	earlier this year
9	1,007	while	832	last year	469	in addition to
10	997	meanwhile	754	for example	466	according to Mr.

Table 2. Most common WSJ reduced *n*-grams

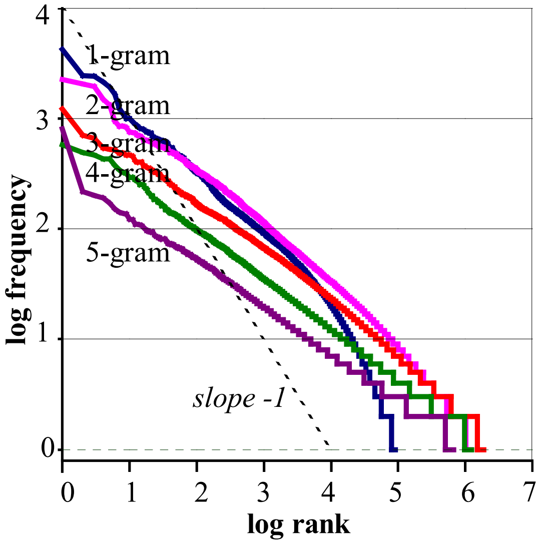

Figure 1. The WSJ reduced *n*-gram Zipf curves

4.2 North American News Text corpus

The NANT reduced *n*-grams are created by the improved algorithm after over 300 hours processing, needing a storage requirement of 100GBytes on a Pentium II 586 of 512MByte RAM.

Their Zipf curves are plotted for reduced unigrams and *n*-grams in Figure 2 showing all the curves are just sloped around [-0.54, -0.5]. The reduced unigrams of NANT still show the 2-slope behavior when it starts with slope –0.54 and then drop with slope nearly –2 at the end of the curve. We have found that the traditional *n*-grams also show this behaviour, with an initial slope of –1 changing to –2 for large ranks (Ha and Smith, 2004.)

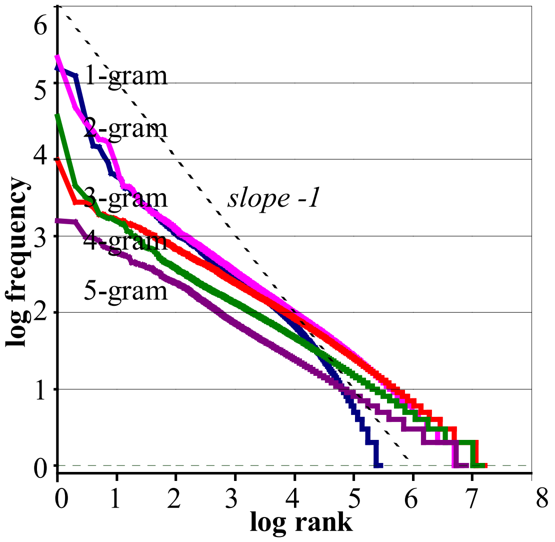

Figure 2. The NANT reduced *n*-gram Zipf curves

4.3 Mandarin News compound words

The Zipf curves are plotted for the smaller Chinese TREC reduced unigrams and *n*-grams were shown by Ha et al. (2005.)

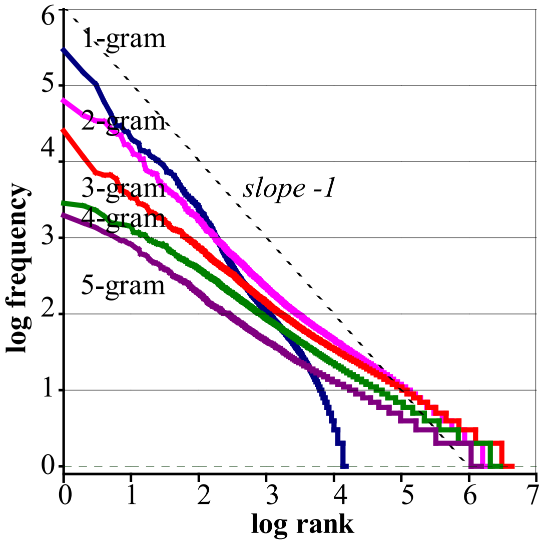

Figure 3. Mandarin reduced *n*-gram Zipf curves

The Mandarin News reduced word *n*-grams were created in 120 hours, using 20GB of disk space. The Zipf curves are plotted in Figure 3 showing that the unigram curve now has a larger slope than –1, it is around –1.2. All the *n*-gram curves are now straighter and more parallel than the traditional *n*-gram curves, have slopes within [-0.67, -0.5].

Usually, Zipf's rank-frequency law with a slope –1 is confirmed by empirical data, but the reduced *n*-grams for English and Chinese shown in Figure 1, Figure 2 and Figure 3 do not confirm it. In fact, various more sophisticated models for frequency distributions have been proposed by Baayen (2001) and Evert (2004.)

5 Perplexity for Reduced *N*-Grams

The reduced *n*-gram approach was used to build a statistical language model based on the weighted average model of O'Boyle, Owens and Smith (1994.) We rewrite this model in formulae (2) and (3)

$$wgt\left(w_j^i\right) = \log\left(f\left(w_j^{i-1}\right)\right) \times 2^{i-j+1} \tag{2}$$

$$P_{WA}\left(w_i \middle| w_{i-N+1}^{i-1}\right) = \frac{wgt(w_i) \times P(w_i) + \sum_{l=1}^{N-1} wgt\left(w_{i-l}^i\right) \times P\left(w_i \middle| w_{i-l}^{i-1}\right)}{\sum_{l=0}^{N-1} wgt\left(w_{i-l}^i\right)} \tag{3}$$

This averages the probabilities of a word w_i following the previous one word, two words, three words, etc. (i.e. making the last word of an *n*-gram.) The averaging uses weights that increase slowly with their frequency and rapidly with the length of *n*-gram. This weighted average model is a variable length model that gives results comparable to the Katz back-off method (1987), but is quicker to use.

The probabilities of all of the sentences w_1^m in a test text are then calculated by the weighted average model

$$P\left(w_1^m\right) = P_{WA}(w_1) P_{WA}(w_2|w_1) \ldots P_{WA}\left(w_m \middle| w_1^{m-1}\right) \tag{4}$$

and an average perplexity of each sentence is evaluated using Equation (5)

$$PP\left(w_1^m\right) = \exp\left(-\frac{1}{L}\sum_{i=1}^{L} Ln\left(P_{WA}\left(w_i \middle| w_1 w_2 \ldots w_{i-1}\right)\right)\right) \tag{5}$$

Ha et al. (2005) already investigated and analysed the main difficulties arising from perplexity calculations for our reduced model: a statistical model problem, an unseen word problem and an unknown word problem. Their solutions are applied in this paper also. Similar problems have been found by other authors, e.g. Martin, Liermann and Ney (1997); Kneser and Ney (1995.)

The perplexity calculations for both the English and Chinese reduced *n*-grams includes statistics on phrase lengths starting with unigrams, bigrams, trigrams…and on up to the longest phrase which occur in the reduced model. The perplexities of the WSJ reduced model are shown in Table 3, North American News Text corpus in Table 4 and Mandarin News words in Table 5.

The nature of the reduced model makes the reporting of results for limited sizes of *n*-grams to be inappropriate, although these are valid for a traditional *n*-gram model. Therefore we show results for several *n*-gram sizes for the traditional model, but only one perplexity for the reduced model.

Unknowns	Tokens	**0**
	Types	**0**
Traditional Model	Unigrams	762.69
	Bigrams	144.33
	Trigrams	**75.36**
	4-grams	60.73
	5-grams	56.85
	6-grams	55.66
	7-grams	55.29
Reduced Model		**70.98**
%Improvement of Reduced Model on baseline Trigrams		**5.81%**
Model size reduction		**14.56**

Table 3. Reduced perplexities for English WSJ

Unknowns	Tokens	**24**
	Types	**23**
Traditional Model	Unigrams	1,442.99
	Bigrams	399.61
	Trigrams	**240.52**
	4-grams	202.59
	5-grams	194.06
	6-grams	191.91
	7-grams	191.23
Reduced Model		**230.46**
%Improvement of Reduced Model on baseline Trigrams		**4.18%**
Model size reduction		**11.01**

Table 4. Reduced perplexities for English NANT

Unknowns	Tokens	**84**
	Types	**26**
Traditional Model	Unigrams	1,620.56
	Bigrams	377.43
	Trigrams	**179.07**
	4-grams	135.69
	5-grams	121.53
	6-grams	114.96
	7-grams	111.69
Reduced Model		**148.71**
%Improvement of Reduced Model on baseline Trigrams		**16.95%**
Model size reduction		**11.28**

Table 5. Reduced perplexities for Mandarin News words

In all three cases the reduced model produces a modest improvement over the traditional 3-gram model, but is not as good as the traditional 4-gram or higher models. However in all three cases the result is obtained with a significant reduction in model size, from a factor of 11 to almost 15 compared to the traditional 7-gram model size.

We did expect a greater improvement in perplexity than we obtained and we believe that a further look at the methods used to solve the difficult problems listed by Ha et al. (2005) (mentioned above) and others mentioned by Ha (2005) might lead to an improvement. Missing word tests are also needed.

6 Conclusions

The conventional *n*-gram language model is limited in terms of its ability to represent extended phrase histories because of the exponential growth in the number of parameters. To overcome this limitation, we have re-investigated the approach of O'Boyle (1993) and created reduced *n*-gram models. Our aim was to try to create an *n*-gram model that used semantically more complete *n*-grams than traditional *n*-grams in the expectation that this might lead to an improvement in language modeling. The improvement in perplexity is modest, but there is a large decrease in model size. So this represents an encouraging step forward, although still very far from the final step in language modelling.

Acknowledgements

The authors would like to thank Dr Ji Ming for his support and the reviewers for their valuable comments.

References

Douglas B. Paul and Janet B. Baker. 1992. The Design for the Wall Street Journal based CSR Corpus. In *Proc. of the DARPA SLS Workshop*, pages 357-361.

Francis J. Smith and Peter O'Boyle. 1992. The *N*-Gram Language Model. *The Cognitive Science of Natural Language Processing Workshop*, pages 51-58. Dublin City University.

George K. Zipf. 1949. *Human Behaviour and the Principle of Least Effort*. Reading, MA: Addison-Wesley Publishing Co.

Harald R. Baayen. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers.

Jianying Hu, William Turin and Michael K. Brown. 1997. Language Modeling using Stochastic Automata with Variable Length Contexts. *Computer Speech and Language*, volume 11, pages 1-16.

Joshua Goodman and Jianfeng Gao. 2000. Language Model Size Reduction By Pruning And Clustering. *ICSLP'00*. Beijing, China.

Kristie Seymore and Ronald Rosenfeld. 1996. Scalable Backoff Language Models. *ICSLP'96*, pages 232-235.

Le Q. Ha and Francis. J. Smith. 2004. Zipf and Type-Token rules for the English and Irish languages. *MIDL workshop*. Paris.

Le Q. Ha, Elvira I. Sicilia-Garcia, Ji Ming and Francis J. Smith. 2002. Extension of Zipf's Law to Words and Phrases. *COLING'02*, volume 1, pages 315-320.

Le Q. Ha, Elvira I. Sicilia-Garcia, Ji Ming and Francis J. Smith. 2003. Extension of Zipf's Law to Word and Character *N*-Grams for English and Chinese. *CLCLP*, 8(1):77-102.

Le Q. Ha, Rowan Seymour, Philip Hanna and Francis J. Smith. 2005. Reduced *N*-Grams for Chinese Evaluation. *CLCLP*, 10(1):19-34.

Manhung Siu and Mari Ostendorf. 2000. Integrating a Context-Dependent Phrase Grammar in the Variable *N*-Gram framework. *ICASSP'00*, volume 3, pages 1643-1646.

Manhung Siu and Mari Ostendorf. 2000. Variable *N*-Grams and Extensions for Conversational Speech Language Modelling. *IEEE Transactions on Speech and Audio Processing*, 8(1):63-75.

Nelson Francis and Henry Kucera. 1964. *Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Department of Linguistics, Brown University, Providence, Rhode Island.

Peter L. O'Boyle. 1993. A study of an *N*-Gram Language Model for Speech Recognition. *PhD thesis.* Queen's University Belfast.

Peter O'Boyle, John McMahon and Francis J. Smith. 1995. Combining a Multi-Level Class Hierarchy with Weighted-Average Function-Based Smoothing. *IEEE Automatic Speech Recognition Workshop*. Snowbird, Utah.

Peter O'Boyle, Marie Owens and Francis J. Smith. 1994. A weighted average *N*-Gram model of natural language. *Computer Speech and Language*, volume 8, pages 337-349.

Ramon Ferrer I. Cancho and Ricard V. Solé. 2002. Two Regimes in the Frequency of Words and the Origin of Complex Lexicons. *Journal of Quantitative Linguistics*, 8(3):165-173.

Reinhard Blasig. 1999. Combination of Words and Word Categories in Varigram Histories. *ICASSP'99*, volume 1, pages 529-532.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-off for *M*-Gram Language Modeling. *ICASSP'95*, volume 1, pages 181-184. Detroit.

Reinhard Kneser. 1996. Statistical Language Modeling Using a Variable Context Length. *ICSLP'96*, volume 1, pages 494-497.

Slava M. Katz. 1987. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume ASSP-35, pages 400-401.

Stefan Evert. 2004. A Simple LNRE Model for Random Character Sequences. In *Proc. of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles*, pages 411-422.

Sven C. Martin, Jörg Liermann and Hermann Ney. 1997. Adaptive Topic-Dependent Language Modelling Using Word-Based Varigrams. *EuroSpeech'97*, volume 3, pages 1447-1450. Rhodes.

Thomas R. Niesler and Phil C. Woodland. 1996. A Variable-Length Category-Based *N*-Gram Language Model. *ICASSP'96*, volume 1, pages 164-167.

Thomas R. Niesler. 1997. *Category-based statistical language models*. St. John's College, University of Cambridge.