Universität des Saarlandes
Fachrichtung 4.7 Allgemeine Linguistik
Studiengang Computerlinguistik

Diplomarbeit

Optimal Design of a Speech Database for Unit Selection Synthesis

Anna Hunecke

Saarbrücken, den 13. September 2007

Durchgeführt beim
Deutschen Forschungszentrum für Künstliche
Intelligenz (DFKI) GmbH,
Saarbrücken
Betreuer: Dr. Marc Schröeder

Danksagung

Ich bedanke mich bei Dr. Marc Schröder und Professor William Barry für die gute Betreuung und Unterstützung während der Erstellung dieser Arbeit, außerdem bei meinem Freund Andreas Maier und meiner Mutter Vera Hunecke für das Korrekturlesen. Weiterhin bedanke ich mich bei meiner Familie und allen Freunden, die mich unterstützt und ermutigt haben.

Erklärung

Hiermit versichere ich, dass ich die Diplomarbeit mit dem Titel "Optimal Design of a Speech Database for Unit Seleciton Synthesis" selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Stellen meiner Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht.

Saarbrücken, den 13. September 2007

Anna Hunecke

Zusammenfassung

Diese Diplomarbeit beschäftigt sich mit der Frage, wie die optimalen Sätze für einen Sprachkorpus für Unit Selection Synthese aus einer großen Satzmenge (dem Textkorpus) ausgewählt werden können. Zur Beantwortung dieser Frage wird ein Algorithmus entwickelt und untersucht, der genau diese Auswahl vornimmt.

Der Algorithmus ist ein gieriger Algorithmus, der bei jeder Iteration den Satz mit dem höchsten Wert auswählt. Dieser basiert auf dem Wert für die einzelnen Laute des Satzes, die durch einen Vektor mit phonetischen und prosodischen Eigenschaften repräsentiert sind. Für die Berechnung eines Lautwerts spielen die zwei Gewichte "Häufigkeit" und "Bedarf" eine große Rolle. Das Häufigkeitsgewicht spiegelt die Häufigkeit der Eigenschaften eines Lautes im Textkorpus wider. Das Bedarfsgewicht dagegen legt fest, wie nötig die Lauteigenschaften im Sprachkorpus gebraucht werden. Wird ein Satz ausgewählt und dessen Laute zum Sprachkorpus hinzugefügt, verringern sich die Bedarfsgewichte für die Eigenschaften aller hinzugefügten Laute.

Der Algorithmus wird zunächst auf einem Englischen Korpus, der sowohl Text als auch die entsprechenden Aufnahmen umfasst, im Rahmen des Wettbewerbs "Blizzard Challenge" getestet. Von den besten Ergebnissen wird eine Satzmenge ausgewählt, aus der eine Stimme für den Wettbewerb gebaut wird.

Für ausführlichere Tests werden zwei deutsche Textkorpora aus den Internetressourcen Projekt Gutenberg (Gutenberg (2007)) und Wikipedia (Wikipedia (2007a)) erstellt. Die Korpora haben eine Größe von 897.096 (Gutenberg) bzw. 2.159.445 Sätzen (Wikipedia).

Vor dem Hintergrund der Erfahrungen in der Blizzard Challenge, und aufgrund der Größe der Korpora, wird der Algorithmus optimiert. Außerdem wird eine Bewertung für die Satzqualität eingeführt, um Sätze, deren phonetische Transkription zweifelhaft ist, auszuschließen.

In einer finalen Testreihe wird zunächst nach den optimalen Einstellungen für die Parameter des Algorithmus gesucht. Weiterhin werden in der Testreihe die Auswirkung unterschiedlicher Korpusgrößen auf das Ergebnis des Algorithmus untersucht. Der letzte Test der Reihe zeigt schließlich, dass eine gründliche Vorauswahl des Textkorpus nötig ist, und wie diese erreicht werden kann.

Die Implementierung des Algorithmus und weitere Programme, die für den Bau eines Sprachkorpus hilfreich sind, werden als Teil des Sprachsynthesesystem OpenMary veröffentlicht (Schröder and Trouvain (2003)).

Contents

List of Figures

List of Tables

1 Introduction

Speech Synthesis has come a long way since the first experiments with artificial speech were made. Two different approaches can be distinguished: the signal-modeling approach and the concatenative approach. The signal-modeling approach creates the speech by modifying or generating an acoustic signal. The concatenative approach glues together chunks of speech from a database of speech recordings, the speech corpus. For both approaches, there are various methods. This thesis concentrates only on unit selection synthesis, a concatenative synthesis method.

Concatenative synthesis used to suffer the problem of lack of computing resources. There was neither enough memory to store large speech copora, nor enough computing power to process these corpora in acceptable time. Therefore, it was restricted to diphone synthesis.

A diphone is a unit used often in speech synthesis: it stretches from the middle of one phone to the middle of the next phone. Diphones are more appropriate units for concatenative synthesis than phones, because with diphones the individual units are glued together in the middle of a phone and not at the phone boundary. The middle of a phone is thought to be the most "stable" part of the phone with the least influence from the surrounding phones. Thus, gluing the units together at this point makes the resulting speech sound smoother.

In diphone synthesis, the speech corpus consists only of diphones, spoken with a very monotone voice. The resulting synthesis is very intelligible, but the monotone intonation makes it sound more like a robot than like a human.

With more computing power, concatenative synthesis moved on to unit selection synthesis, where the corpus consists of real sentences, spoken naturally. The speech units are often diphones, but can also be bigger (phrases) or smaller (halfphones). Units for synthesis are selected according to the two measures "target cost" and "join cost". "Target cost" measures how well a unit matches to the unit specification derived from the text. It is based on the phonetic and prosodic properties of a unit. "Joinc cost" measures the smoothness of the transition of a unit to the neighboring units in the speech signal.

The quality of unit selection synthesis is much higher in regard of naturalness than diphone synthesis. On the downside, this kind of synthesis is more likely to have audible joints between units because of the different intonation and

stress of the units. This is exactly the reason why the corpus consists only of monotone speech in diphone synthesis.

Another problem of unit selection synthesis is the size of the speech corpus: Although large corpora are now possible, from the computing power point of view, they are still problematic: Firstly, the recording of a large speech corpus takes a lot of time and money: For good quality, a professional speaker has to be hired and paid, and professional equipment and a recording room are needed. Furthermore, during recording, every utterance has to be checked for correct pronunciation, and re-recorded, if necessary. This makes the recording process very slow.

Also, the speech corpus has to be divided into units. This is called labeling, and can be done automatically - but the quality is, of course, much better, when the unit boundaries are hand-corrected - which also costs money. And, of course, the speech synthesis program must be capable of handling a large corpus efficiently.

For these reasons, the size of a speech corpus is an important factor to consider when building a speech synthesis voice. While it is, on the one hand, determined by the available resources - that is money -, on the other hand, the purpose the voice is built for, what it should be capable of to synthesize later on, also plays a role. If the voice will only be used in a limited domain - for example, train travel information or speaking clock - the set of sentences that will be synthesized with it might be limited. If the domain of the voice is general, however, and the voice is expected to be able to say everything with a reasonable quality, this raises the question of coverage.

What is coverage? Coverage defines what kind of acoustic realizations are present in a speech corpus. The more different acoustic realizations, the better the synthesis, since there are more choices for a synthesis unit and the chance is higher that there are units that match well. Coverage can be regarded on different levels, like phone or diphone coverage. Full diphone coverage, for example, means that the speech corpus contains all diphones that are needed.

This is the desired state, because missing diphones make the synthesis sound worse: A missing diphone has to be constructed from two other diphones and results in a potentially audible joint at the phone boundary.

But in unit selection, it is not enough to have full diphone coverage. To reduce the possibility of audible joints, it is also desirable to have the same diphone several times in different prosodic realizations. For example, a diphone can be stressed or not, can be high-pitched or lower-pitched. However, constructing a speech corpus with full coverage of prosodic variations is a difficult task, because the number of prosodic variations is very high.

But how is a speech corpus constructed anyway? The basic idea is to take

a large set of sentences from which to select a subset for recording. To avoid confusion, the set of sentences from which sentences are selected is referred to as "text corpus" or "text database", and the set of sentences selected for recording as "speech corpus" or "speech database". The selection of sentences is done automatically with the help of a selection algorithm.

The goal of this thesis is to develop and implement such an algorithm, to be able to construct a speech corpus with good coverage. The performance of the algorithm is assessed by applying it to an English text corpus and two German text corpora.

Chapter 2 presents previous work in the field of speech corpus construction. It gives an overview over the findings of other researchers on the nature of unit distribution and presents selection algorithms used in other approaches. Against this background, the main features of the selection algorithm are presented in the last section of this chapter.

In chapter 3, the algorithm is described in more detail. Apart from the actual selection process, the definitions of units and coverages used by the algorithm are given. The chapter also describes the implementation of the algorithm.

In the following chapter, chapter 4, the algorithm is applied on an already recorded English speech database, which is used as text database in this context. From the resulting speech corpus, a synthesis voice is built and evaluated.

Based on these tests, the algorithm is refined. This is described in chapter 5. The chapter also documents the optimizations made to the algorithm for handling larger text corpora and presents a method to remove unreliable sentences from a text corpus.

After that, chapter 6 describes the building process of two German text corpora.

These corpora are used to assess the performance of the final implementation of the algorithm. These tests and their results are described in chapter 7.

Chapter 8 describes the sentence selection tool that implements the algorithm developed in this thesis and is distributed as open source code together with Mary (Schröder and Trouvain (2003)). Mary is an open-source Text-to-Speech System developed at the DFKI (German Research Center for Artificial Intelligence) in Saarbrücken.

Finally, chapter 9 gives a summary of the findings in the thesis.

2 Related Work

2.1 Unit Distribution

In this section, the nature of unit distribution in a text database is investigated. In section 2.1.1, the work of van Santen (1997) is presented, to give a first idea of what kind of distribution can be expected in a text database. After that, in section 2.1.2, the peculiarities of such a distribution are further specified, based on Möbius (2003). Section 2.1.3 describes Andersen and Hoequist (2003)'s approach to handle the distribution.

2.1.1 The nature of unit distribution

Van Santen (1997) analyses the combinatorial distribution of units in text databases. To this end, he transcribes 250,000 sentences from the Associated Press Newswire Corpus into diphones. For each diphone, he creates a contextual vector. This vector contains additional information about the diphone, such as accent and position in the utterance. 222,678 different types of context vectors are found in the sentences.

Based on this, the author measures the probability that the domain can be covered adequately. For this, he defines two sets of sentences: the training and the test set. The vector types in the training set are taken as the definition of the acoustic inventory of the domain. The test set is a set of random sentences from the same domain. The question is: how big does the training set have to be, to ensure a high probability that all vectors from the test set are in the training set?

Van Santen (1997) finds that the probability is only 0.03 when the training set contains 25,000 different vector types. To reach a probability of 0.75, a training set with more than 150,000 different vector types is needed. Given that these values are obtained with the test and the training set being from the same domain, van Santen predicts that coverage will be worse if the two sets are from different domains. He concludes that a speech database with a good coverage of vector types has to be very large - too large to be feasible.

Van Santen (1997) also investigates the differences of triphone and vocabulary distributions in different corpora. His results show that the differences

are huge: For two different text corpora (169,328 personal names and 347,857 sentences from Associated Press Newswire without proper names), 47.5% of the triphone types occur in one corpus but not in the other. Also, for a number of different corpora, even the two most similar corpora (Associated Press Newswire from 1990 and 1991) only show a correlation of 0.71. Thus, the author concludes that "In all domains investigated, it proved impossible to obtain either complete coverage or at least very high values of the coverage index [...] using training databases of a practically viable size."(van Santen (1997), section 5).

2.1.2 LNRE distributions

The coverage problem is also described in Möbius (2003). Möbius characterizes the distribution of acoustic units in a database as a LNRE (Large Number of Rare Events) distribution. LNRE distributions have the characteristic that some events occur often, while the majority of events occurs rarely. Since the number of rare events is high, the probability that a rare event occurs is high. Thus, just ignoring the rare events because they are rare is likely to lead to severe coverage problems. Möbius suggests "... to increase the coverage of a speech database by carefully definining the linguistic and phonetic criteria that the database should meet" (Möbius (2003), section 4).

2.1.3 Handling LNRE distributions

Andersen and Hoequist (2003) argue that the LNRE distribution of concatenation units can be handled by defining a hierarchy of prosodic features. For example, they state that stress has more influence on the realization of a phone than position of a phone in a sentence, and thus should be treated as more important. The division of the features into important and less important features is reflected by dividing the definition of the coverage of a speech corpus into "target coverage" and "full coverage". "Target coverage" contains the combinations of those features that are most important to have in the speech database, whereas "full coverage" encloses all possible combinations of all features. The authors state that "... with careful planning, target coverage is not only possible but also feasible with a surprisingly small set of properly crafted sentences" (Andersen and Hoequist (2003), section 2).

An analysis of diphones in three different Danish corpora leads the authors to the conclusion that the diphones that only occur between words add to some extent to the problem of LNRE: When only the diphones that occur inside words are taken into account, the number of possible diphones is decreased, which at the same time reduces the number of rare diphones. Therefore,

Andersen and Hoequist (2003) suggest that the fact that a diphone is either part of a syllable or the bound of two syllables should be used in the hierarchy of features.

2.2 Selection Algorithms

Various algorithms have been used to select sentences for a speech database. The algorithm used the most is the greedy algorithm. The paper of van Santen and Buchsbaum (1997), described in section 2.2.1, gives a good overview of the basic greedy algorithm and some of its most common variants. Section 2.2.2 discusses François and Boëffard (2002), in which the greedy algorithm is compared to two other selection algorithms. Section 2.2.3 presents the work of Bozkurt et al. (2003). In this paper, a modified greedy algorithm is described. Finally, in section 2.2.4, Black and Lenzo (2001) is described, in which acoustical measures are used as the basis for sentence selection.

2.2.1 Basic greedy algorithm

Van Santen and Buchsbaum (1997) give a good introduction into the so-called greedy algorithm, the algorithm used most frequently in selection tasks. The basic idea is that, at each iteration, the algorithm selects the sentence which maximizes some sort of selection criterion, until a stop criterion is reached. The algorithm is called greedy because of the strategy of always selecting the local optimum.

In more detail: the algorithm starts with an empty set, the cover set, which has to be filled by the algorithm. The sentences in the filled cover set then form the speech corpus. Furthermore, there is a set of sentences (the text corpus) and another set containing the list of diphones for each sentence. The goal is to fill the cover set with sentences from the text corpus so that all diphone types in the diphone set occur at least once. At each step of the greedy algorithm, the sentence with the best count of unseen diphone types is removed from the text corpus and added to the cover set. The algorithm continues until N sentences are in the cover set.

This basic algorithm can be varied to take into account the frequency of the diphones: diphones that occur more frequently in the sentence set are seen as more valuable than diphones that are rare, because they will most probably be needed more often in the synthesis than the rare ones.

Another variant to inverse the frequencies of the units, so that rare units are considered more valuable than frequent units. The idea behind this is that the more frequent units will end up in the cover set anyway: Since they

are so frequent, they are likely to occur in the sentences containing the rare units. This measure better accounts for the LNRE distribution of units. In van Santen and Buchsbaum (1997), inversing the frequency reduces the number of selected sentences needed to reach full coverage by up to 10%.

However, as already mentioned in chapter 1, units in unit selection synthesis are not just plain old diphones, but can also be vectors containing phonetic and prosodic information. Of course, this representation greatly enlarges the number of possible units, since every possible combination of the vector values (called "features") has to be taken into account.

Van Santen and Buchsbaum (1997) propose to reduce this number by identifying those features that have an effect on each other. As an example, the authors give the effect of position in a phrase on the duration of a vowel. The combinations of those features that do not interact are ruled out, reducing the number of feature vectors that have to be covered.

This approach is used in Shih and Ao (1994) to build a corpus for their duration study of Mandarin Chinese. The text database used consists of 15,630 sentences with 1,385,451 units. The authors take into account 11 different features, as well as the two features unit identity and tone identity. Every unit is represented by 11 feature vectors, consisting of the latter two features and one of the other 11 features. As a result,8,233 different feature vectors are identified in the database. They could be covered in a speech corpus of just 424 sentences using the greedy selection algorithm.

2.2.2 Alternatives to the Greedy algorithm

François and Boëffard (2002) analyze three different heuristics of selecting sentences for a speech database: a greedy algorithm, a spitting algorithm and a pair-exchange algorithm.

The greedy algorithm has been described above. In contrast to that, the spitting algorithm works exactly the opposite way: Initially, all sentences of the text corpus are in the cover set. At each step, the most useless sentence is removed. This goes on until a stopping criterion is met.

In the pair-exchange algorithm the cover set is initially filled with an arbitrary set of sentences from the text database. At each step, a sentence from the cover set is compared with a sentence from the text database. The sentence with the higher score is stored in the cover set and the sentence with the lower score is put into the text database. This process goes on until it is stopped. In contrast to the other algorithms, the number of sentences in the cover set does not change in the course of the algorithm.

François and Boëffard (2002) compare the different algorithms by running them each over a text corpus of 3000 sentences. Different variations of the

algorithms with different selection criteria for the sentences are used. The selection criteria are based on the number of useful and useless unit types and instances in a sentence: A useful unit type is one that is needed in the cover set, whereas a useless unit type is not needed. Unit instances are useful if their type is useful and there are still instances of the type needed in the cover set, otherwise they are useless. The sentence length is also taken into account.

For comparing the results of the tests, François and Boëffard measure the number of unit instances, average sentence length and number of selected sentences of the resulting speech corpora. The authors state that the organization and balance of selection criteria can greatly influence the result and the length of the selected sentences: From an average sentence length of 52.4 instances in the text corpus, the average length of sentences varies from 17.6 to 136.3 instances in the different speech corpora.

From the results of the tests, the authors conclude that the performance of greedy and spitting algorithm is equally good. They also state that both guarantee full coverage, whereas the pair-exchange algorithm does not and is too time consuming.

François and Boëffard (2002) furthermore state that the spitting algorithm is costly in comparison to the greedy algorithm. The reason for this is that only a small part of the text database is selected in the end. Because the spitting algorithm starts with the full text database in cover, a high number of sentences has to be removed, and a lot of iterations are needed. For the greedy algorithm to come to the same result, the number of iterations needed corresponds to the number of selected sentences.

However, the authors conclude that the spitting algorithm is still useful, since applying it to the result of the greedy algorithm enhances the result: "The spitting algorithm removes up to 18.6% sentences and 10.8% instances. It does it rapidly, for us its utility is blatant" (François and Boëffard (2002), section 6.2).

2.2.3 Modifying the greedy algorithm

In Bozkurt et al. (2003), a modified version of the greedy algorithm is proposed: while the other approaches worked with a concrete definition of which units are needed in the speech corpus, in this approach the algorithm is aimed to maximize the number of different units.

As in the other approaches, units are feature vectors with phonetic and prosodic features. Selection is based on the sentence score: for each sentence, the score is the normalized sum of the scores for the units of the sentence.

Calculation of the unit score is based on the instances of the current unit in the cover set: For each of them, a so-called "MatchScore" is computed. The

MatchScore reflects the similarity of the unit instance in the cover set to the current unit: the lower the MatchScore, the more similar the units. Since the units in the resulting speech corpus have to be as different as possible, higher MatchScores are better. The lowest MatchScore of all instances of the current unit in the cover set is taken as the score of the current unit. If there are no instances of the unit in the cover set yet, the unit score is 1.

The MatchScore is computed as follows[1]:

$$MatchScore = 1 - \sum_{n=1}^{N} w(n) * F(n)$$

$w(n)$ is the weight of feature n and $F(n) \in \{0, 1\}$ is the matching score of feature n. $F(n) = 1$ if the values for that feature of the unit in the cover set and the current unit match, and 0 otherwise. If there is more than one value for a feature and the values of the units match, $F(n)$ is 1 divided by the number of values.

Bozkurt et al. (2003) state that the advantage of the approach is that the number of features used is practically not restricted, since there is no definition of what features are wanted in the resulting speech corpus. Thus, the whole problem of unit distribution is ignored.

The authors test the performance of their modified greedy algorithm on two text databases containing 2500 sentences each. For comparison, the standard greedy algorithm optimizing for diphone coverage is also applied to these text databases.

In the tests, the feature vectors only consist of the diphone feature and the feature "phonetic context", that is, the neighboring phones of the diphone. From the results, the authors conclude that the modified greedy algorithm selects a phonetically richer corpus than the normal greedy algorithm.

Alas, the modified method has a high computational load. Therefore, Bozkurt et. al propose to combine it with the standard greedy selection: First, corpus size is reduced with the normal greedy method, and then the modified method is applied to the results.

They use this approach to build a Turkish speech corpus: From a text corpus of 115,000 sentences, 20,000 sentences are selected with the standard greedy algorithm. Then the modified method is used to select 5,000 and 2,500 sentences, respectively, out of this reduced text corpus.

[1] from Bozkurt et al. (2003)

2.2.4 Greedy selection based on acoustic models

A completely different usage of the greedy selection algorithm for the creation of a speech corpus is proposed by Black and Lenzo (2001). Their idea is to base the selection of sentences on acoustical measures.

Their approach works as follows: The units of an existing speech corpus are clustered according to acoustic similarity. Then, all sentences of a text database are synthesized with the voice built from the speech corpus, and the number of uses of each unit cluster is counted. Finally, the greedy algorithm is used to select sentences for a speech corpus from the text database. The selection is based on the frequency of use of the unit clusters.

For the first step, Black and Lenzo (2001) build a cluster unit selection voice from an existing speech corpus. This is a unit selection voice, but the synthesis units are clustered according to acoustic similarity. The unit clusters are stored in the leaves of a decision tree, which is used at runtime to select a cluster. Only the units in the selected cluster are taken into account for the synthesis. The questions in the tree are phonetic features.

The authors then proceed in synthesizing their whole text database. The database consists of 19 novels from Project Gutenberg[2]. During synthesis, the number of times a unit cluster is used is counted.

The third step consists of applying the greedy selection algorithm to the text database. The selection criterion for the algorithm is the sum of the score of the units of a sentence. The unit score is obtained by traversing the decision tree containing the clusters for each unit. The score is the frequency of the cluster reached. If there is already an instance of the unit in the cover set, the unit gets the score 0.

Black and Lenzo (2001) apply the algorithm to their text database two times. At the first pass, 221 sentences are selected, and, at the second pass, 146 sentences. Three different speech corpora are created from these two sets: one for every set and one combining the two sets. For assessing the quality of the speech corpora, all selected sentences are synthesized with the cluster unit selection voice, and three voices are created.

The voices were evaluated via listening tests. The test sentences are from various domains: from novels, from another speech database, from a communicator application and from a scientific paper. Not surprisingly, the voice with the biggest corpus performs best. Also, the test sentences from the story domain get a higher rating better than those from the other domains.

The authors point out that their method relies heavily on the speaker of the original speech database from which the clusters are built, and that the selected

[2]Project Gutenberg is an online collection of out-of-copyright prose. More on this in chapter 6.

sentences will be different for each speaker. This is supported by further tests with a different speech database. This is also the advantage of this method: sentence selection does not rely on general text processing methods, but is optimized for individual speakers. The downside is the computational load and the effort needed.

2.3 Conclusion

The work presented in this chapter gives an overview over the different dimensions to take into account for the design of a speech corpus. From this background, general considerations about the structure of the selection algorithm that is to be developed can be made.

As shown in section 2.1, the definition of the units is an important factor. On the one hand, the definition should account for as many different prosodic variations of diphones as possible. On the other hand, the LNRE distribution of the units restricts the number of prosodic properties that can be used. In the unit definition used in this thesis, explained in detail in section 3.1, there is only one feature with six different values to cover the most common prosodic variations.

The presentation of different selection algorithms in section 2.2 has shown that the greedy algorithm is most appropriate for the task at hand. For this algorithm, the selection criterion is the most important parameter to define. The presented approaches based the selection on a frequency measure (van Santen and Buchsbaum (1997), Black and Lenzo (2001)), as well as on the units that are already in the cover (François and Boëffard (2002), Bozkurt et al. (2003)). The selection algorithm presented in this thesis takes both measures into account.

3 Preliminary Algorithm

This chapter gives a first outline of the algorithm that was developed throughout the work on this thesis. The algorithm, a variant of the greedy algorithm, is described in detail in section 3.2. Before that, section 3.1 states what kinds of units were used as the basis of the selection algorithm. Section 3.3 describes the coverage measures that were applied to measure the quality of the results. In section 3.4, the details of the implementation of the algorithm are explained. Section 3.5 summarizes the algorithm characteristics.

While the ultimate goal is to run the algorithm on a large corpus, the version of the algorithm described in this chapter was used only on a smaller corpus. Based on the results, amendments were made as described in section 5. The tests and results on this preliminary version are described in section 4.

3.1 The units

Units are defined as vectors consisting of four features. For each phone, there is one feature vector. The four features are phonetic identity, phonetic identity of the next phone, phone class of the next phone and prosodic property of the current phone.

As the preliminary algorithm is tested on English, the English phone set of Mary is used as phone definition. This phone set defines 41 different phones plus the zero phone. The zero phone is used to mark the end of a sentence. Thus, there are $41 * 42 = 1722$[1] different possible diphones.

The concept of phone classes was introduced to reduce the number of possible diphones. The idea behind this is that the transitions in the middle of two diphones are similar if the second parts of the diphones are similar phones.

For example, the transitions from a vowel to an alveolar consonant will be the same or very similar, no matter which alveolar consonant it is. However, they will be distinct from the transitions of that vowel to a velar consonant. For the consonants, the place of articulation is more important for the transitions than the manner. The same is true for the vowels: the dimensions open-closed and front-back are more important than rounding. In this manner, 21 phone

[1] There are only 41 instead of 42 possible phones for the first half of a diphone, because the first half can not be the zero phone.

classes were defined, reducing the number of possible diphones from 1722 to $41 * 21 = 861$. The phone classes for English are listed in appendix A.

For the prosodic properties of a phone, six different properties were defined: unstressed, stressed, pre-nuclear accent, nuclear accent, phrase final high and phrase final low. These features represent the most important prosodic variations. Stress/no stress is determined by lexical stress. The accents and phrase final tones are computed on the basis of ToBI predictions.

3.2 The algorithm

An algorithm using greedy methods is used for sentence selection. Three major parameters interacting with each other influence the selection of sentences:

- **Coverage definition:** The definition of coverage fixes what kinds of units are wanted in the final set. Two different settings were used for the algorithm.

  The first setting, *simpleDiphones*, defines the cover as all combinations of the three features phonetic identity, phonetic identity of the next phone and prosodic property of the current phone. In other words, the *simpleDiphones* setting considers all diphones and their prosodic variations.

  For the second setting, *clusteredDiphones*, all combinations of the three features phonetic identity, phone class of the next phone and prosodic property of the current phone are considered. This way, the cover is defined as all combinations of *clustered diphones* and their prosodic variations.

- **Sentence score:** For each unit token, a certain score determines how "useful" the token is for the selected set. For each sentence, the score is the sum of the scores of the units in this sentence divided by the number of units.

  The score of a unit token is basically the product of two different weights: frequency weight and wanted weight. The frequency weight reflects the frequency of the unit type in the text corpus. The wanted weight determnines how much a unit type is "wanted" in the speech corpus. The intuition is that if there is already a token of a particular type in the cover, the wanted weight should be lower than if there is no instance of the type in the cover yet.

  These two weights are computed for a unit token on all levels of the cover: on the phone level (feature phone), the diphone level (features

nextPhone and nextPhoneClass, respectively) and on the prosody level (feature prosody). The sum of the three products is the score of the token.

Both weights can be set to a number of different settings.

For the frequency weight, three settings are possible: 1 (which means no consideration of frequency), relative frequency (which gives a preference for the more common units) or 1- relative frequency (1minus for short, gives a preference to the rarest units).

For the wanted weight the variation of settings is high. On every level, the wanted weight can be set to a different value. Thus, a high setting on the phone level and lower settings on the two other levels would render unit tokens that are uncovered phones more useful than unit tokens that are uncovered prosodic variations of covered diphones. This way, the settings can be optimized for phone, diphone or prosodic coverage.

The preference for the different levels is not so much controlled by how high or low a value is for a level, but rather the relation between the values are important: Setting the weight to 10 on the phone level and to 1 on the diphone level makes new phones ten times more useful than new diphones. The same effect can be achieved by setting the weight on the phone level to 100 and on the diphone level to 10.

An additional dimension is added by the setting for the decrease of the wanted weight: Each time a unit token is selected for the cover set, the wanted weight for this unit type is divided by a certain number, to reflect the fact that we already have this type and do not necessarily want another instance of it. The higher this number, the less useful it is to add unit types that are already represented in the cover.

Yet another dimension of the sentence score is the sentence length: a sentence that is longer or shorter than a certain threshold value is given the score 0, which prevents it from being selected. The reason for restricting the sentence length is that too long or too short sentences are difficult to record, and more likely to be of worse quality.

- **Selection function:** The selection function determines the basis on which the next sentence is chosen. There a two possibilities: A simple function is to select the sentence which has the most new units to add to the coverage. In order not to favor long sentences, the number of new units has to be divided by the number of units in the sentence. A more enhanced selection function bases its selection on the sentence score.

In the preliminary algorithm, two selection functions are used: one based on the highest sentence score, and one based on both sentence score and number of new units.

The selection function keeps selecting the best sentence until a stop criterion is reached. In the first tests, the stop criterion is the total duration of the recordings of the selected sentences. The stop criterion is refined later, in chapter 5.

3.3 Measuring the corpus distribution

Corpus distribution or coverage indicates how many different kind of units are in a corpus. Four main measures for coverage will be used throughout this thesis: simple diphone, simple prosody, clustered diphone and clustered prosody coverage.

Simple diphones coverage measures how many different combinations of phone and nextPhone are in the corpus. The value is calculated by dividing the number of combinations in the cover by the number of possible combinations. Multiplying this value with one hundred gives the percentage.

Similarly, clustered diphones coverage measures the distribution of the combinations of phone and nextPhoneClass.

Simple prosody and clustered prosody coverage are one step more specific. They measure the number of different combinations of phone, nextPhone and prosody and phone, nextPhoneClass and prosody, respectively.

For all four coverage measures it holds true that it is unrealistic that all combinations will be in the text corpus, simply because not all combinations might occur in the language. Therefore, none of the values is expected to reach 1.

3.4 Implementation

The algorithm was implemented in Java and was fit to be integrated into the Mary system.

To save resources, most data was represented in low level data types like arrays and not in objects. Units, for example, are arrays consisting of four bytes.

The sole exception is the cover set: it is represented as a tree. There is a cover set for the simple diphone coverage and one for the clustered diphone coverage. Figure 3.1 is a schematic illustration of the simple diphone cover set.

It has three levels: phone, next phone and prosody. It works like a decision tree for the units: On the phone level, for example, the byte value representing the phone of a unit is at the same time the index of the daughter that represents the phone. At the other levels, it works the same way. Therefore, the feature vector representing the units is at the same time the path down the coverage tree. Every node has its frequency and wanted weight. They are set

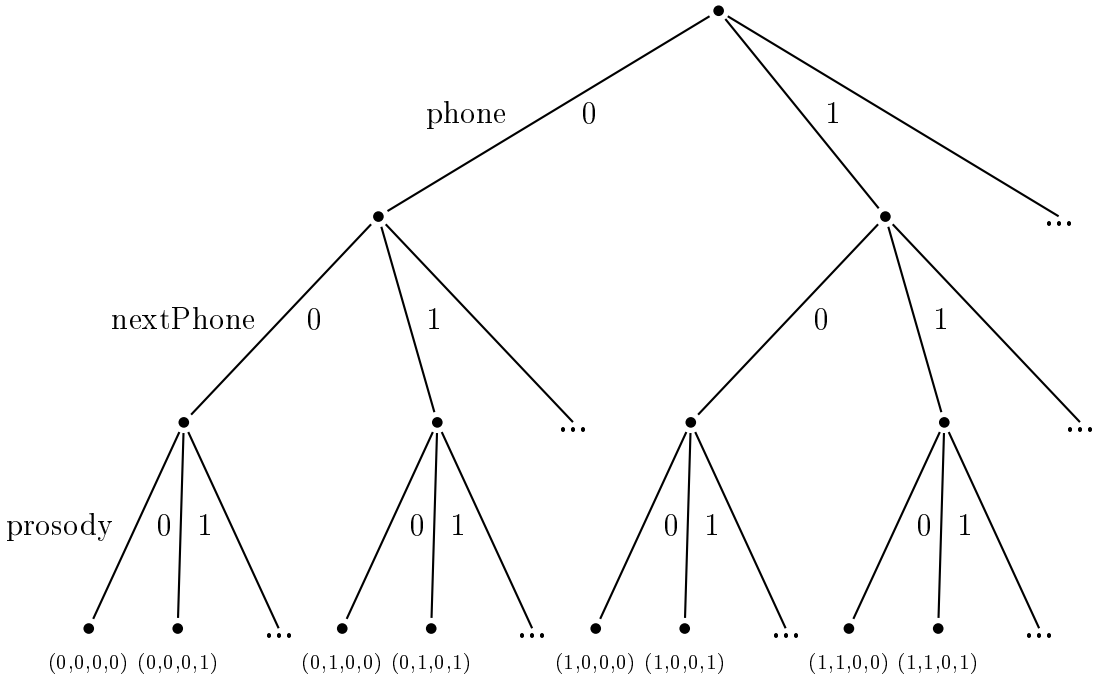

Figure 3.1: Schematic illustration of the coverage set for simple diphones. The vectors in the leaves illustrate which feature vectors the leaves can represent. As this is the tree for simple diphones, the next phone class feature (the third value) is not queried in the tree.

to their initial values at the start of the algorithm. This works as follows: On initialization, the algorithm gets a list of file names as well as the name of a text file where the settings are defined. Each of the files in the list contains the units for one sentence. All files are read and the appropriate units are stored in memory. Also, for every unit, the path down the tree defined by this unit is followed until a leaf is reached. In every leaf, the number of units whose paths lead to that leaf is stored. This is taken as the total frequency and is the basis of the frequency measures. For the nodes higher up, the total frequency is the sum of the frequencies of their children.

At runtime, calculating the score for one unit is just a matter of walking down the tree and summing up the scores of the nodes passed by.

3.5 Summary

In this chapter the algorithm used for the initial experiments has been described. It is basically a greedy algorithm.

The units are feature vectors consisting of the four features phonetic identity, phonetic identity of the next phone, phone class of the next phone and prosody. For selection, however, only three features are taken into account. This is either phone, nextPhone and prosody (simpleDiphones) or phone, nextPhoneClass and prosody (clusteredDiphones).

Selection is based on sentence score and number of new units, respectively. Sentence score is the normalized sum of unit score, which is defined as the sum of products of frequency weights and wanted weights for each of the three features of a unit. The wanted weight defines how much a unit type is wanted in the speech corpus and is decreased each time an instance of that unit type is added to the cover set. The frequency weight reflects the frequency of a unit in the text corpus.

The coverage of the selected sentences is measured with four different measures: As there are two definitions of diphones, namely *simple* (phone + next phone) and *clustered* (phone + next phone class), their coverage has to be assessed separately. Additionally, the coverage of the prosodic variations of the two kinds of diphones is measured.

4 Testing the preliminary algorithm

For the tests of the algorithm described in section 3, the data of the Blizzard Challenge 2007 was used.

The Blizzard Challenge is a speech synthesis competition: The participating teams all get the same speech database from which they have to build a voice for their system and synthesize a set of sentences. The synthesized sentences of all teams are evaluated with listening tests by the Blizzard Challenge organizers. Since all teams get the same data, the results of the evaluation are a good indicator of how good the system sounds in comparison to other systems. The Mary team took part in the challenge for the second time.

This year, three voices were to be created from the given speech database: one from the full database (Voice A), one from the Arctic subset of the database (Voice B) and one from a subset of the database that could be chosen by the teams individually (Voice C). The task for Voice C was an excellent opportunity to test the algorithm.

While the use of the Blizzard Challenge data restricts the selection of text, the advantage is that all text is already recorded. Because of this, a voice can be built and tested directly after selection.

Section 4.1 details the properties of the Blizzard speech database. In the course of selecting sentences to build voice C, the algorithm was applied to the database 288 times with different settings. This is described in section 4.2. In section 4.3, the results of these tests are discussed.

Several voices were built from the sentences selected by the algorithms. One of them was selected to be submitted to the Blizzard Challenge as voice C. The performance of this voice in the challenge is discussed in section 4.4. Section 4.5 summarizes the findings in this chapter.

4.1 Statistics of the corpus and expectations

The Blizzard database consists of 6579 sentences (477 minutes of speech, approximately 8 hours). The speaker is male and the language American English. Although it is - strictly speaking - a speech database, the Blizzard database

	Blizzard	Arctic
Number of sentences	5879	1030
Average sentence length	48.16 units	35.36 units
Maximum sentence length	183 units	73 units
Minimum sentence length	4 units	7 units
phone coverage	100.00%	100.00%
simple diphone coverage	81.65%	77.12%
simple prosody coverage	53.50%	34.71%
clustered diphone coverage	86.30%	81.77%
clustered prosody coverage	60.65%	41.54%

Table 4.1: Statistics of the Blizzard Corpus and the Arctic Corpus

is referred to as text database from now on to avoid confusion. To get an idea of what can maximally be expected of the algorithm, the statistics of the database were computed. The results are shown in table 4.1.

The table shows that the text corpus does not have complete coverage. As for simple diphone coverage, the Blizzard corpus covers 81.65% of the possible diphones and 53.50% of all prosodic variations of the diphones. For clustered diphones, the numbers are slightly better: 86.30% of the possible diphones and 60.65% of the possible prosodic variations of the diphones are covered.

The third column of the table shows the distribution of the Arctic corpus, a subset of 1030 sentences of the Blizzard corpus. As can be expected, the distribution of this smaller corpus is worse: Only 77.12% of the possible diphones and 34.71% of the prosodic variations for these diphones are realized. Again, the numbers are better for clustered diphones. Here, 81.77% of the possible diphones and 41.54% of the prosodic variations are covered.

As can be seen from the table, only 5879 out of the 6579 sentences of the full Blizzard corpus are used. This is because the problematic sentences are sorted out. “Problematic”, in this case, refers to sentences containing words that are not in the dictionary. Most of them are foreign words, mostly Japanese or Spanish. For these sentences, the accuracy of the phonetic transcription can not be guaranteed.

The distributions of the Blizzard and the Arctic corpus are the maximum and minimum of what the algorithm should achieve. The expectation is, that the better the distribution of the resulting speech corpus, the better the sound of the voice. Therefore, it can be expected that a voice built from a speech corpus which has a total duration (in terms of speech recordings) of the same length as the Arctic corpus, but with a better distribution, sounds better than

the voice build from the Arctic corpus.

The tests were also expected to answer a question concerning the division into simple and clustered diphones: good coverage for simple diphones also guarantees good coverage for clustered diphones. The question is if this is also true the other way round: Does good clustered diphone coverage also result in good coverage for simple diphones? The question is answered in section 4.3.1.

4.2 Testing setup

As the algorithm has a lot of parameters to tweak, the amount of tests conducted was very high.

Only the stop parameter of the selection function was not altered, as it was defined by the Blizzard Challenge rules. The total length of the recordings of the selected sentences should not be more than 2914 seconds. This is exactly the length of the Arctic corpus.

The following list gives an overview of the different settings used to test the selection algorithm:

- **Selection functions:**
  - maximum usefulness selection function: the function selects the sentence that is most useful
  - maximum usefulness and maximum number of new units selection function: the function selects the sentence that is most useful and adds the most new units to the cover set.

- **coverage settings:**
  - simple diphones: units are prosodic variations of simple diphones ($41 * 42 * 6 = 10332$ possible different units)
  - clustered diphones: units are prosodic variations of clustered diphones ($41 * 21 * 6 = 5166$ possible different units)

- **frequency weight settings:**
  - do not consider frequency
  - relative frequency
  - 1minus (1- relative frequency)

- **wantedWeight settings:**
  - phone level 100; diphone level 10; prosody level 1: new phones are ten times more useful than new diphones; and new diphones are ten times more useful than new prosodic variations
  - phone level 1; diphone level 1; prosody level 1: new phones, new diphones and new prosodic variations are all equally useful

- divide wanted weight by 10000, 1000, 100, 10 when an instance of a unit is selected: the higher this number, the more useful uncovered units compared to units already seen.

- **sentence length settings:**
  - do not consider sentence length
  - minimum/maximum sentence length 10/150 units
  - minimum/maximum sentence length 10/170 units

The number of tests conducted amounts to 288.

4.3 Results

4.3.1 Parameter settings

The performance of the algorithm depends heavily on the settings of the parameters. The effects of the parameter's settings interact with each other, so that there is no ultimate setting for one parameter. Nevertheless, some general trends can be observed.

As to the definition of the units: simple diphones as units generally maximize both simple diphone and clustered diphone coverage, whereas the clustered diphone units do not maximize the coverage for simple diphones.

Frequency weights: Generally, both 1minus frequency and no frequency lead to a better distribution than relative frequency. The former two often lead to very similar results. One explanation for this could be the unit distribution: Since the frequency is close to 0 for most units, the 1minus frequency is close to 1 for most units, too - which is also the value they will get when the frequency is not considered. The question is: is the 1minus frequency useful at all, if similar results can be achieved without regarding the frequency?

The two different settings of the wanted weights optimize for different distribution properties: Setting the wanted weights to 100/10/1 maximizes the diphone coverage, whereas setting them to 1/1/1 maximizes the coverage of prosodic variations. However, for the prosodic variations, 100/10/1 is also a good measure. It would have been interesting to test the setting 1/10/100, but this is left over for the next round of tests.

The setting of the value by which the wanted weights are divided has very little influence. Often, the differences are marginal. This indicates that the settings are too similar. For 1000 and 10000, the results are often the same. The reason for this is probably that when a wanted weight is divided by one of those two numbers, it is practically out of count, since it is in any case

	v1	v2	v6
Number of sentences	931	921	836
Average sentence length	35.53 units	35.79 units	39.39 units
Maximum sentence length	113 units	109 units	113 units
Minimum sentence length	4 units	4 units	4 units
phone coverage	100.00%	100.00%	100.00%
simple diphone coverage	81.64%	79.15%	76.66%
simple prosody coverage	47.08%	47.53%	44.66%
clustered diphone coverage	86.30%	83.86%	86.30%
clustered prosody coverage	55.17%	55.59%	59.39%

Table 4.2: Distributions of the three selected voices

significantly smaller than the non-decreased wanted weights. Dividing the wanted weight by 10 often leads to worse results than the other three settings.

In general, the selection function considering only the usefulness of the sentence and not the number of new units leads to a better distribution. The selection function using usefulness and number of new units sometimes leads to a better prosodic coverage, but most of the time the results are worse for this selection function. Also, with the latter function, often the results are the same for 1minus frequency and no frequency.

The settings for the restriction of the sentence length do not make much of an impact. In general, setting no restriction on sentence length leads to better results. The two different restrictions on the sentence length often lead to the same results - not surprising, because they are very similar.

4.3.2 Resulting Voices

Based on the test results, the resulting speech corpora with the best distributions are chosen to build test voices. It turned out that there is no setting which maximizes all four different coverage measures. Therefore, three different voices are built: v1, v2 and v6. Their coverage distributions are shown in table 4.2.

The speech corpus of voice v1 maximizes both simple diphone and clustered diphone coverage; those two coverages are the same as for the text corpus. The speech corpus of voice v2 maximizes only the prosody coverage for simple diphones. The maximum of the text corpus (53.50%) can not be reached, but the percentage is higher than for the Arctic corpus. The speech corpus of voice v6 maximizes the coverage of clustered diphones and clustered prosody. Again,

	v1	v2	v6
selection function	usefulness	units+usefulness	usefulness
consider sentence length	false	false	false
wanted weights	100/10/1	100/10/1	1/1/1
divide wanted weights by	100	10000;1000	100
frequency settings	1minus	1minus;none	1minus

Table 4.3: The algorithm settings for the three voices

the coverage of the clustered diphones is the same as in the text corpus, but the clustered prosody coverage is lower than in the full corpus(60.07%).

4.3.3 Best settings

The settings which lead to the best distributions are shown in table 4.3. For all three voices, the sentence length is not considered. This indicates that some important phones are in the sentences that are ruled out when the sentence length is restricted.

For the speech corpora of voices v1 and v6, the selection function considering only the usefulness of a sentence is used, while the speech corpus of v2 is selected with the selection function considering usefulness and number of new units. This selection function weakens the influence of the settings: The distribution of the speech corpus of v2 can be reached both by setting the frequency weight to "1minus" and by not considering the frequency at all. It also does not matter whether the wanted weight is divided by 1000 or by 10000.

The wanted weight settings for the speech corpora of voices v1 and v6 reflect the general trend that 100/10/1 gives a preference for diphones and 1/1/1 a preference for prosody. The speech corpus of voice v2 is also unusual in this respect.

4.3.4 Quality of the selected sentences

The distribution statistics only show how many different phones are in the selected sentences, but are no measure of the quality of the selected sentences. The sentence length can be an indicator for this: Very long sentences are considered worse, because recording longer sentences is more error prone than recording short sentences. In the speech corpus of voice v2, the majority of the sentences has a length between 10 and 30 units, while in the speech corpus of v6, the length of most of the sentences varies between 10 and 40 units. The

	v6	Arctic	full
MOS	3.0	2.8	3.2
similarity	2.9	3.3	3.7
WER	0.200	0.203	0.103

Table 4.4: Results of the three voices from the DFKI-team in the Blizzard Challenge 2007

sentence length in the speech corpus of v1 lies in between.

Since the sentences are transcribed automatically, it is very probable that there are transcription errors in the selected sentences. When looking at the sentences in the speech corpora, some are very obvious. For example, the sentence "uh-uh." is transcribed as "ʌʌ". Unfortunately, this is a rare diphone, which is why the sentence is included in the speech corpora of all three voices.

4.3.5 Informal listening tests

To decide which of the voices is submitted to the Blizzard Challenge, informal listening tests are performed. For each voice, 20 test sentences are synthesized: 10 from the news domain and 10 from the story domain (namely, the first 10 sentences of "Alice's Adventures in Wonderland"). The differences between the voices are minimal.

In the end, voice v6 is chosen, because it has the most natural intonation and sounds smoother than the others. In comparison, voice v1 sounds choppy and unnatural. The sound of voice v2 is in between the other two. For all voices, some obvious synthesis errors occur in the test sentences, but, from a subjective point of view, the amount of errors is roughly the same for every voice.

4.4 Results of the voice in the Blizzard Challenge

Table 4.4 shows the results of voice v6 in the Blizzard Challenge in comparison with the results for the full voice (built from the whole Blizzard Corpus) and the Arctic voice.

The following measures of synthesis quality are used[1]:

[1]Definitions are based on Wikipedia (2007c), Wikipedia (2006) and Wikipedia (2007b)

"MOS", or Mean Opinion Score, denotes the mean of the listeners opinion on the synthesized sentences. The values range from 1 to 5, with 1 being the lowest score and 5 the highest. There were two sets of tests in two different domains in which the MOS was assessed: conversational domain and news domain. In the tests, the subjects had to rate the naturalness of the sentences.

"Similarity" means the similarity to the original speaker. This was rated in tests where the subjects had four audio recordings of the original speaker and one synthesized utterance. The subjects had to decide if the synthesized utterance sounded as if it was the same person as in the recordings. The scale for this rating ranges also from 1(totally different person) to 5(exactly the same person). The values shown in the table are the mean of the subject's opinions.

Finally, "WER" (Word Error Rate) assesses how intelligible the synthetic speech is. WER is computed on the basis of the Levenshtein distance. Originally, the Levenshtein distance is a measure for the distance between two words: It is defined by the minimum number of substitutions, deletions and insertions of characters that have to be conducted to get from one word to another. Similarly, the Word Error Rate measures the distance of two sentences: Word Error Rate for a sentence is the minimum number of substitutions, deletions and insertions of words that is needed to get to the "right" sentence (called target sentence). The values of WER lie between 0 (no errors) and 1 (everything wrong).

To compute the Word Error Rate, semantically unpredictable sentences (SUS) were used. This means, that the sentences are grammatically correct, but do not make any sense. A typical example is "The migratory laggards froze the replete biz.". The subjects had the task of listening to a synthesized SUS and write down the words they could understand. The WER was computed on the basis of the subject's transcriptions.

The problem with this measure is that the test was also attended by non-native speakers of English. It is unclear how much general problems of understanding played into their judgments, especially since there are a number of difficult and unusual words in the sentences. Therefore, the values shown in the table are only the mean of the results for the subjects that are native speakers.

The table shows that, with respect to the mean opinion score, v6 is actually better than the Arctic voice but worse than the full voice. This would back up the assumption that a better distribution leads to better performance.

However, regarding the other two measures, v6 does not perform better than Arctic: The word error rates of Arctic and v6 are approximately similar, and worse than for the full corpus. Also, v6 is judged less similar to the original speaker than the other two voices.

The results for the word error rate indicate that the coverage of the two smaller corpora was not enough to ensure an intelligibility as high as that of the full corpus. But the mean opinion score indicates that this is not necessarily the crucial factor for listeners.

4.5 Summary

This chapter has described how the algorithm presented in chapter 3 was used to select sentences for a voice for the Blizzard Challenge. To this end, the algorithm was applied to the provided text database 288 times with different settings. From the resulting speech corpora, the three speech corpora with the best coverages were used to build three voices. The best voice was identified with the help of informal listening tests and submitted to the Blizzard Challenge. In the challenge, the voice performed better (in terms of mean opinion score) than the voice built from the Arctic subset of the voice database, and worse than the voice built from the full database.

5 Refining the algorithm

The Blizzard Challenge was a good opportunity to test the algorithm and to enhance it, before it is applied to the German corpora. This chapter describes the conclusions drawn from these tests and how they affect the algorithm.

First of all, another frequency measure is introduced in section 5.1. Also, the stop criterion is redefined, as described in section 5.2, since the Blizzard stop criterion can not be used for the final tests. Section 5.3 explains why one of the two selection functions is dropped. Section 5.4 describes the changes made to the implementation of the algorithm. In addition to the enhancement of the algorithm, a check for sentence credibility, described in section 5.5, is introduced before the processing.

5.1 Frequency

As seen in the tests, the frequency measures 1minus and no frequency show a very similar performance. The original goal of promoting the rare units can not really be achieved by the 1minus measure. Rather, it makes all units more or less equal.

Therefore, inverse frequency is introduced as an additional frequency measure. Inverse frequency is computed by dividing 1 by the relative frequency. This gives the desired boost to the selection of rare units, since the small differences between the relative frequency values are made significantly wider.

One example to illustrate this: Rare unit a has a relative frequency of 0.0005, whereas common unit b has a relative frequency of 0.05. The 1minus frequency is 0.9995 for unit a and 0.95 for unit b. The inverse frequency is 2000 for unit a and only 20 for unit b. Thus, with inverse frequency unit a is a hundred times more useful than unit b, while with 1minus frequency, unit a is only about 1.052 times more useful than unit b.

5.2 Stop criterion

Little attention has been paid to the stop criterion so far. The most obvious criterion is to stop when the maximum coverage which can be achieved with

the given corpus is reached. Since we have four coverage measures, it is only appropriate to have four different stop criteria; one for every coverage measure.

A restriction of the number of sentences is also a useful stop criterion, since the recording of the data will be restricted to some number of sentences.

All stop criteria are used in the final tests.

5.3 Selection function

As pointed out in sections 4.3.1 and 4.3.3, the selection function using both usefulness and number of new units is difficult to handle. Since we do not know how many new units a sentence will contain, the number of new units is a factor hard to control. It weakens the influence of the settings and thus makes the selection unpredictable.

Therefore, this selection function is dropped altogether and not used in the final algorithm.

5.4 Implementation

The basics of the implementation do not have to be changed. Yet, some alterations have to be made to account for the huge increase of corpus size.

Firstly, the representation and handling of features is optimized to allow for more sentences to be kept in memory. Instead of using a separate byte array for each unit, all units of a sentence are stored in one array. Also, the reading of the feature files is optimized for speed by using buffered methods.

But there is still a limit to the number of sentences that can be kept in memory. Therefore, the implementation is altered, so that the sentences are either read into memory when the program starts, or stay on hard disk throughout the run of the program and are read each time they are needed. If they stay on disk, this slows down the selection considerably.

For example, if the algorithm is used to select 500 out of 5000 sentences, it takes about 15 minutes when the sentences stay on disk. When the sentences are read into memory, it takes only 45 seconds[1].

The initialization of the algorithm takes a long time, because the coverage of the text corpus has to be computed. Thus, after the first initialization, all the coverage data computed in the initialization is saved in a binary file on disk. In subsequent runs of the algorithm, this initialization file is read in - this speeds up the initialization phase. If the sentences stay on disk, the acceleration of

[1] The computation was conducted on a 64-bit PC with Pentium 4 3.2 GHz CPU and 2 GB RAM

the initialization is very obvious. If they are read into memory, the speedup is not so noticeable, because each sentence has still to be loaded into memory.

5.5 Unknown words

In order for the algorithm to work and produce meaningful results, one has to make sure that the features of the sentences are correct. Sentences with potentially incorrect features should not be taken into account at all. Therefore, they should be removed from the text corpus before the algorithm is applied to it.

To get reliable features, the phonetic transcriptions of the sentences have to be correct. In a Text-to-Speech Synthesis System such as Mary, words are transcribed by looking them up in a lexicon. The lexicon used for the German transcription in this thesis is the Mary lexicon for German. It contains 137,565 roots of words, which can be expanded to 615,449 full words. Also, it contains about 3,000 proper names.

Although the lexicon is quite large, it is clear that the lexicon does not contain every possible word. If a word is not in the lexicon, the transcription is computed with some heuristic. In Mary, compound analysis and Anglicism analysis is used to obtain a transcription. If this fails, so called letter-to-sound rules are used. These are general rules that model the German pronunciation. The transcriptions produced by them are generally not very reliable.

English words are transcribed with the English module, which has only the options to use the English lexicon or English letter-to-sound rules for transcription.

The credibility of a sentence is based on the method used to obtain its phonetic transcription. For this, first of all, transcription tags are introduced. The tags indicate with what kind of method a word was transcribed.

There are seven different tags for German text:

- **lexicon:** The word is in the lexicon. This is the most credible transcription.
- **userdict:** The word is in the lexicon specified by the user. Also credible.
- **phonemiseDenglish:** The word is an Anglicism and is transcribed by the Denglish (=Deutsch+English) module. This implies either cut off of inflections or compound analysis. Is not as credible as the lexicon.
- **compound:** The word is a German compound and was transcribed with the help of a compound analysis. Not as credible as the lexicon.

lexicon	100
userdict	100
phonemisedenglish	50
compound	50
foreign_en	50
rules	10
preprocessed	10
nothing	100

Table 5.1: The credibility weights for the different tags

- **foreign_en:** The word is an English word. Less credible than lexicon, since the word can be transcribed either using the English lexicon or the English letter-to-sound rules and it is unknown which method was used.
- **rules:** The word is transcribed using letter-to-sound rules. Least credible method.
- **preprocessed:** The word was transcribed by the preprocessor. Also not very credible.

Additionally, the tag **nothing** is defined for words that have no transcriptions, such as punctuation. This tag is, of course, very credible.

For each tag, a weight that reflects the credibility of transcriptions tagged with it is set. Each unit gets the tag of the word that it is part of. The credibility of the whole sentence is the sum of the credibility per unit divided by the number of units.

Table 5.1 shows the setting of the weights. With these settings, a sentence will get the score 100 if it is a hundred percent credible. If the credibility score is below a certain threshold, the sentence is considered not credible.

Section 6.2 describes how this method is applied in the creation of the German corpora.

5.6 Summary

In this chapter, the modifications applied to the algorithm first described in chapter 3 are presented. They incorporate the results of the algorithm in the Blizzard Challenge and the larger amount of data to be processed later on. A new frequency criterion - inverse frequency - is introduced, as well as new stop criteria. These are based on the maximum coverage to be achieved and on the

maximum number of sentences to select, respectively. The selection function taking into account the number of new units is discarded. The implementation of the algorithm is optimized to cope with a larger set of sentences. A check of sentence credibility is introduced to reduce the possibility of transcription errors.

6 Building two German text corpora

In order to obtain more robust results about the performance of the algorithm, it has to be applied to a bigger text corpus than just 5879 sentences. To this end, two German text corpora are constructed: one covers the story domain and the other one the dictionary domain.

Section 6.1 describes the considerations taken in the design of the text corpora. The next section, 6.2, gives a description of the creation process of the corpora. Finally, in section 6.3, the distribution statistics of the text corpora are shown.

6.1 Design issues

General design issues have to be settled prior to the building of the corpus.

At first, the language: Although the first tests were conducted on an English corpus, the language of the corpora is German. The reason for this is that they will be used in research at the DFKI later on.

The next consideration regards the domain of the corpora: the original idea was to build several corpora for different domains in order to be able to compare test results for different domains. However, this intention is hampered by practical issues: the corpora have to be freely available and also free for potential future re-distribution.

From these considerations, two German text sources in the Internet were selected: Gutenberg (Gutenberg (2007)) and Wikipedia (Wikipedia (2007a)).

As mentioned in section 2.2.4, Gutenberg is a website containing out-of-copyright books. The aim of Project Gutenberg is to distribute them freely. As of April 17th 2007, 377 so-called ebooks were available in German on the Gutenberg homepage. The inventory consists of stories, plays, poems and specialized books. Most of them are more than 70 years old, making the language of the content old-fashioned.

Wikipedia is an open encyclopedia in the Internet. Everyone can contribute, write or modify entries. The content is released under the GNU Free Documentation License (GFDL). As of 3rd of July 2007, there were 605,629 German

entries. The style of the texts is formal German.

Because the language of the text corpora changes to German, the phone set changes also. There are 61 phones plus the zero phone in the Mary phoneset for German. Of these, 7 (ø, e, i, o, u, y, œ̃) are unused, because they only appear in foreign words, and can be represented by their tensed versions (øː, eː, iː, oː, uː, yː, œ̃ː). So, in the end there are 54 phones plus the zero phone.

For the clustered diphones, phone classes were defined on the same basis as the English phone classes, as described in section 3.1. For German, 27 phone classes are defined. They are listed in appendix B.

6.2 Corpus building

This section describes the creation of the text corpora. The general procedure is as follows: First, the text is divided into sentences. Then, unreliable sentences are sorted out and, for each sentence, the phonetic and prosodic features are computed. The implementation of the necessary steps is done in Java and Perl. Mary is used throughout the implementation.

6.2.1 Gutenberg corpus

First of all, the ebooks were downloaded from the Gutenberg homepage with GNU wget. This is a program for retrieving files via HTTP and similar protocols. The ebooks are available in several formats. For the current task, only the plain text versions are used.

In the next step, the texts were split into sentences. This was done with the help of Mary: The text was processed paragraph by paragraph to get an xml representation of the text, including tags for sentence start and end. Each of the sentences was then broken down into individual phone units for which the five features phonetic identity, phonetic identity of next unit, phone class of next unit, prosodic type and credibility of the transcription were computed. The feature computation was also done with Mary.

An extra processing step was introduced at the point were a word is looked up in the lexicon. As Gutenberg contains some very old texts, the spelling is sometimes very old-fashioned: “ß” instead of “ss” and “th” instead of “t” are the most common manifestations. Therefore, if a word was not found in the lexicon, some simple normalization steps were taken to remove the old-fashioned spelling. Then the word was looked up again. If that second lookup was also unsuccessful, the original version of the word was passed on to the other transcription modules.

The final step in the creation of the corpus was to determine which sentences are credible. Thus, for each sentence the credibility score was computed as described in section 5.5. The credibility threshold was set to 90. This removed about one tenth of the sentences in the corpus:

There were quite a lot of sentences in Latin, Greek, French, Italian, Spanish, English or Old-English which did not pass the credibility check.

Also, a lot of sentences containing unusual names were defined as incredible. The names were mostly Greek (*Aristoteles*, *Sokrates*), Latin (*Donatus*, *Virgil*) or French (*Voltaire*, *Corneille*).

Furthermore, many old-fashioned phrases were deemed not credible: For example, the sentence *Den eilften*[1] *Abend(mittewochs, den 6. Mai) ward Miß Sara Sampson aufgeführet.* would today be written like this: *Am elften Abend(Mittwochs, den 6. Mai) wurde Miss Sara Sampson aufgeführt.* (*On the eleventh evening (Wednesday, 6th of May), (the play) Miss Sara Sampson was performed.*).

Tables of contents and other listings were also often deemed incredible.

The examples show that the credibility check has successfully picked out unsuitable sentences. The total size of the corpus shrank from 978,273 to 897,096 sentences.

6.2.2 Wikipedia corpus

The Wikipedia encyclopedia is too large to be downloaded with wget. Instead, a compiled version in the form of one large xml file was downloaded. The size of the unzipped file was approximately 3.2 GB. It contained all articles, templates, image descriptions, and primary meta-pages that were part of the German version of the Wikipedia when the file was archived (27th of April, 2007).

The xml file was then split into individual articles. This step included the removal of the xml structure. Each article was saved in its own file. This resulted in 605,629 individual files.

In the next step, the articles had to be converted to plain text. This step was the most difficult step, since the articles contained not only Wikipedia commands, but also html commands. The removal was done with regular expressions.

As done with the Gutenberg corpus, the articles were then split into sentences and their features were computed with Mary. This was done in one step.

[1] This is actually a spelling mistake; it should be "elften"

Also in this step, incredible sentences were removed. The threshold was set to 87. The following list gives an overview of the types of sentences that were removed.

- sentences with unknown compounds: German is known for its high number of compounds: Practically every word can be combined with another in German. In particular, the formal, written German used in the Wikipedia contains a lot of compounds.

  An example for a sentence with problematic compounds: *Lohbrügge liegt am nördlichen Rand des Elbe Urstromtales im Regenschatten der Harburger Berge.* (*Lohbrügge lies at the northern edge of the glacial valley of the Elbe in the rain shadow of the Harburg mountains.*): *Urstromtales* is analyzed as compound, but *Regenschatten* is wrongly analyzed as Anglicism ('ʁeː-gəns-tʃɛ-tən).

- sentences with proper names: In the Wikipedia, there are a lot of articles about films, actors, medicine topics and locations containing foreign words. The situation is not so problematic when English words are involved: *Top Gun ist ein Spielfilm aus dem Jahre 1986.*(*Top Gun is a feature film from 1986.*): *top* is defined in the user dictionary, and *gun* is analyzed as an English word; both are transcribed correctly.

  But the transcription of Icelandic names, for example, is of course unreliable: *Die alternativen Namensvorschläge Hnjúkabyggð, Tröllaskagabyggð und Ægisbyggð konnten sich beim Wahlvolk nicht durchsetzen.* (*The alternative suggestions Hnjúkabyggð, Tröllaskagabyggð und Ægisbyggð could not become accepted by the voters.*). All three place names are transcribed using rules: German letter-to-sound rules applied to Icelandic words are very unlikely to produce the right transcription.

- sentences with unknown German proper names: Also, there are a lot of articles about German locations: *In Bitburg Erdorf zweigt die Nims Sauertalbahn nach Bitburg(Stadt) ab, welche ursprünglich über Irrel bis nach Igel führte.* (*In Bitburg Erdorf the Nims Sauertalbahn branches off to Bitburg(City), which originally led to Igel over Irrel.*). While *Bitburg* and *Igel* are actually in the lexicon, *Erdorf* and *Irrel* are analyzed as compounds (ʔɛʁ-'dɔɐf and ʁɐ-ʁeː-ʔɛl) and *Nims* is actually analyzed as English word (nɪmz).

- sentences in foreign languages: Sometimes there are whole sentences in a foreign language in the Wikipedia, for example Nauru (*Nauru bwiema, ngabena ma auwe.*) or Chinese (*Bù yuàn zuò núlì de rènmen!*).

	Gutenberg	Wikipedia	Wikipedia897096
Number of sentences	897096	2159445	897096
Average sentence length	88.35 units	117.97 units	118.90 units
Maximum sentence length	9450 units	18919 units	10766 units
Minimum sentence length	1 unit	1 unit	1 unit
phone coverage	100.00%	100.00%	100.00%
simple diphone coverage	73.84%	80.84%	78.75%
simple prosody coverage	54.21%	63.81%	60.51%
clustered diphone coverage	75.03%	79.56%	77.92%
clustered prosody coverage	58.06%	65.08%	62.63%

Table 6.1: Distribution of the Gutenberg corpus, the Wikipedia corpus and the first 897096 sentences of the Wikipedia corpus

- sentences with Wikipedia formatting: Some Wikipedia commands slipped through the format removal in the previous step, leading to meaningless sentences. Examples: *{ |class = prettytable !* or *thumb |right |Zentralasien mit Seidenstrasse Kysylkum (usbek.)*

The removal of unwanted sentences took place during the construction of the corpus. The number of rejected sentences was quite high: from 3,527,902 sentences, 1,368,453 were deemed unreliable, leading to a total number of 2,159,445 sentences.

The whole building process for the Wikipedia corpus took several weeks more than the creation of the Gutenberg corpus. In the end, only approximately one third of the Wikipedia (the first 204,521 of the 605,629 articles) was used due to time constraints.

6.3 Corpora distribution

Table 6.1 shows the distributions of the two text corpora. Additionally, the distribution of the first 897,096 sentences of the Wikipedia corpus is displayed, as this sub corpus will be used in the tests in chapter 7.

The Wikipedia corpus is more than twice as large as the Gutenberg corpus. It also contains longer sentences and has a higher diphone and prosody coverage than the Gutenberg corpus. However, the differences in coverage of about 4-10% are not as large as one would expect given the difference in size. Interestingly, the coverage of the first 897,096 sentences of Wikipedia is still better than the coverage of the Gutenberg corpus.

It catches the eye that the diphone and prosody coverages are still far away from 100%, even in the larger corpus. As it is likely that not all possible combinations occur in German, it can be assumed that the coverage of the Wikipedia corpus is near the maximum of what can be reached, since the corpus is so large.

Figure 6.1 illustrates the distribution of simple diphones in the corpora. The y-axis shows the frequency of a diphone type, and the x-axis the number of types. As could be expected, the distributions show the characteristics of LNRE distributions. In comparison, the Wikipedia corpus has a more even distribution. This means, the Gutenberg corpus has a higher number of diphones that occur only once.

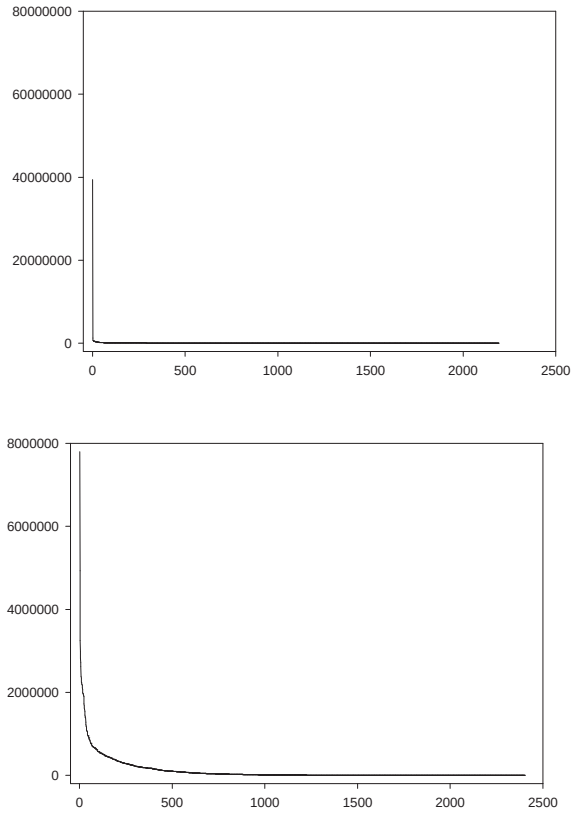

Figure 6.1: Diphone distribution in the Gutenberg corpus (top) and Wikipedia corpus (bottom). Diphones are sorted according to frequency.

7 Final tests and results

This chapter describes the tests conducted on the text corpora described in the previous section. Three main goals are pursued:

Firstly, the goal was to find out the best settings for the algorithm - the settings that yield the best coverage. The appropriate tests and their results are described in section 7.1.

Then, the influence of the corpus size was examined to answer the question of how big a database has to be to ensure good coverage. The intuition is that coverage grows with database size. The tests conducted to answer that question are discussed in section 7.2.

Finally, an example synthesis script was produced and evaluated to show what the algorithm can achieve. "Synthesis script" in this context refers to the actual list of sentences in a speech corpus. This is described in section 7.3.

7.1 Finding the best settings

The first goal of the tests is to find out the best settings for the four different coverage measures simple diphone, simple prosody, clustered diphone and clustered prosody coverage. To achieve this goal, the algorithm was applied repeatedly to the same text corpus with different settings.

In order to test all combinations of settings, the algorithm had to be conducted 96[1] times.

Because of this high number, only a small text corpus, consisting of the first 5,000 sentences of the Gutenberg corpus, was used. The results were verified by repeating the tests on the first 5,000 sentences of the Wikipedia corpus, thus raising the number of tests conducted to $96 * 2 = 192$.

At each pass, 500 sentences were selected and their coverage was measured. Section 7.1.1 describes which settings were used, and section 7.1.2 discusses the results of the tests.

[1] 2 diphone definitions $*$ 4 frequency settings $*$ 4 wanted weight settings $*$ 3 wanted weight decrease settings $= 96$

7.1.1 Settings

In the following, the different values that were assigned to the settings are listed.

First of all, there are the two settings for the unit definition: Simple Diphones and Clustered Diphones.

Then there are the four different settings for the frequency weight: none, normal, 1minus and inverse. With setting "none", the sentence score depends entirely on the wanted weight. Setting "normal" gives a preference for the most frequent units, while "1minus" gives a slight preference for rare units. Setting "inverse" gives a very strong preference for rare units.

For the wanted weight, also four different settings are chosen: 25/5/1 (phone level 25, diphone level 5, prosody level 1), 1/5/1, 1/5/25 and 1/1/1.

With setting 25/5/1, new phones are 5 times more valuable than new diphones and new diphones 5 times more valuable than new prosodic variations. This setting is expected to result in a good diphone coverage.

Setting 1/5/1 is a variation of 25/5/1, and is expected to perform similarly. The reasoning behind this expectation is the same as the one behind the preference for rare units: there are "only" 54 phones (opposed to, for example, $54 * 55 = 2970$[2] simple diphones and $54 * 55 * 6 = 17820$ prosodic variations of simple diphones), and thus every phone has a high likelihood in comparison with simple diphones or simple prosodic variations. Because of this, it is very likely that the resulting speech corpus will contain all different phones, although they are not explicitly selected.

Setting 1/5/25 gives a preference to new prosodic variations over diphones and new diphones over new phones. This setting is expected to lead to a high prosodic coverage.

Setting 1/1/1 makes new phones, new diphones and new prosodic variations equally valuable.

Another parameter is the number by which the wanted weight is divided when a given unit is selected. In the tests, three different settings are tried out: 2, 5 and 1000.

Dividing the wanted weight by 2 means that, after the first instance of a unit type is found, the balance of the weights still remains, since both the weights 1 and 5 and the weights 5 and 25 are different by factor 5. This setting is expected to benefit the selection of several instances of one unit type.

Setting 5 is the setting that evens the weights of the levels. For example, together with wanted weight setting 25/5/1, once a phone is selected, its wanted weight value is reduced to 5, and it is then as valuable as a new diphone.

[2]Like the English diphones, the zero phone can not be the first half of a German diphone; therefore there are only 54 possibilities for the first half of a diphone

	#	units	wanted weight	frequency weight	divide wanted weight by
simple diphones	1	SD	25/5/1, 1/5/1	inverse	1000, 5
	2	CD	1/1/1	none, 1minus	2
	3		1/5/25	normal	
simple prosody	1	SD	no clear best settings	inverse	1000, 5
	2	CD		none, 1minus	2
	3			normal	
clustered diphones	1	CD	25/5/1, 1/5/1	inverse	1000, 5
	2	SD	1/1/1	none, 1minus	2
	3		1/5/25	normal	
clustered prosody	1	CD	no clear best settings	inverse	1000, 5
	2	SD		none, 1minus	2
	3			normal	

Table 7.1: Overview over the best settings for the algorithm, subdivided into the four coverage measures. "SD" means "simple diphones" and "CD" means "clustered diphones"

Setting 1000 practically reduces the wanted weight of a unit type to zero once an instance of a unit type is added to the cover. This way, only unseen unit types can contribute to the sentence score.

The sentence length parameter was not used in these tests, since it reduces the size of the database by an unknown number of sentences.

7.1.2 Results and Discussion

The detailed results of the tests are listed in appendix C. Table 7.1 gives an overview sorted according to the coverage measures.

In general, the use of simple diphone units leads to better results than clustered diphone units for both simple diphone coverage and simple prosody coverage. Vice versa, clustered diphone units yield better results for clustered diphone coverage and clustered prosody coverage.

As for the different frequency measures, inverse frequency is the clear winner. With inverse frequency, the best results are achieved for all four coverage measures. On the other side, relative frequency is the loser, as most of the bad results are produced with this measure. As in the first tests described in section 4.3.1, 1minus frequency and no frequency perform relatively similar. Their performance is somewhere in the middle between inverse and normal

frequency.

In comparison with the frequency, the impact of the settings of the wanted weight is less clear.

The results for 25/5/1 and 1/5/1 are very similar; with inverse frequency, results for both settings even are almost always the same. These two weights seem to be the best settings for the wanted weight: the algorithm often produces the best results with these two wanted weight settings and inverse frequency weight. Also, when these weights are combined with the other frequency measures, often the results are the best results that can be achieved with that particular frequency weight.

But at least for both prosody coverage measures, the other two wanted weight settings are also important. Both settings sometimes outperform 25/5/1 and 1/5/1. Performance of setting 1/1/1 is generally better than 1/5/25, but there are a few exceptions.

The number by which the wanted weight is divided has the least influence on the result of the selection algorithm. Often dividing by 1000 and 5 leads to better results than dividing by 2.

As for the individual coverage measures, the best settings are not always the same for the Gutenberg and the Wikipedia corpus. For simple diphone and clustered diphone coverage, the results are very similar for both corpora. The inverse frequency measure is apparently the best setting, but only in combination with the wanted weights 25/5/1 and 1/5/1, respectively. Settings none and 1minus are also relatively close to the top, also in combination with 25/5/1 and 1/5/1. Then follows wanted weight setting 1/1/1. The combination of frequency measure normal and wanted weight setting 1/5/25 is the worst setting for diphone coverage.

The differences between the two corpora manifest themselves best in the simple prosody coverage measure. For the Gutenberg corpus, wanted weight setting 1/1/1 performs best, followed by 25/5/1 and 1/5/1, and finally by 1/5/25. In the results for the Wikipedia corpus, 1/5/1 and 25/5/1 are ranked top, followed by 1/5/25 and then 1/1/1.

Obviously, the prosodic coverage is not that dependent on the wanted weight. The reason for this could be the high number of prosodic variations: Because they are so many, it is likely that there are new prosodic variations in every sentence. To explicitly benefit the selection of new prosodic variations with the help of the wanted weight seems not to be absolutely necessary.

In contrast to this, the influence of the frequency weight is reflected quite clearly by the results: all top-ranked results use the inverse frequency measure. Then comes a mix of none and 1minus, and the low ranks are all occupied by the normal frequency measure.

This ranking of frequency measures is not quite as clear in the results for clustered prosody coverage. Similar to the results for the diphone coverage measures, the inverse frequency measure is best, but not in combination with all wanted weight settings. Settings "none" and "1minus" are also relatively good.

However, regarding the wanted weight settings, the picture is as vague as in the results for simple diphone coverage. 1/1/1 and 1/5/25 seem to be the best settings for the Gutenberg corpus, while for the Wikipedia corpus, the other two settings also play a role.

To sum up, the best settings to use for obtaining a good coverage are "inverse" for the frequency setting, a wanted weight setting which sets a preference to the "next phone" feature and "next phone class" feature, respectively, and a high number by which the wanted weight is divided.

7.2 The influence of corpus size

As seen in section 6.3, the Wikipedia corpus, more than twice the size of the Gutenberg corpus, exceeds the coverage of the Gutenberg corpus only by 4-10%. To investigate the relation between text corpus size and speech corpus coverage, several tests were conducted with text corpora of different sizes. First of all, it was tested how many sentences are needed to maximize the possible coverage. This was tested for all four coverage measures. For each measure, the appropriate stop criterion was used: the algorithm stopped once the maximum coverage was reached for the given coverage measure.

Related to this is the question if the coverage of speech corpora of the same size correlates to the size of the text corpora from which they were selected. The expectation is that this is the case, since a larger text database is likely to have more variations.

7.2.1 Setup

The tests were conducted on both German text corpora. The algorithm was applied on the first 10,000, 20,000, 50,000, 100,000, 200,000 and 500,000 sentences of each text corpus. Additionally, the tests were performed on all 897,096 sentences of the Gutenberg corpus and on the first 897,096 sentences of the Wikipedia corpus. They were not conducted on the full Wikipedia corpus, because of time constraints: it is not possible to store the whole Wikipedia corpus in memory, therefore the sentences would have to be read from hard disk - which makes the program much slower.

The settings that performed best in the previous tests were used to ensure a good coverage. These are the settings inverse frequency, wanted weight 25/5/1 and wanted weight divided by 1000. They are the best settings for the diphone coverage and among the best settings for the prosody coverage in the previous tests. Since the differences between the top results in the tests are marginal for the prosody coverage measures, the described settings are chosen over the real best settings for better comparison.

The algorithm was applied to each of the text databases five times with different stop criteria: In the first pass, the algorithm stopped after the maximum simple diphone coverage was reached. In the second pass, the stop criterion was maximum simple prosody coverage. This was repeated with the clustered diphone and the clustered prosody coverage stop criterion. In the fifth pass, the algorithm stopped after selecting 1000 sentences.

In the passes where simple diphone or simple prosody coverage was the stop criterion, and in the pass stopping after 1000 sentences, the units were simple diphones. In the two passes using clustered coverage stop criteria, the units were clustered diphones.

All in all, 70 passes of the algorithm were conducted.

7.2.2 Results and Discussion

Figures 7.1 and 7.2 illustrate the results of the tests.

Figure 7.1 shows the coverage of the resulting speech corpora when the algorithm is stopped after 1000 iterations. The x-axis denotes the size of the text database from which the sentences are selected. The y-axis shows the percentage of coverage.

Not surprisingly, the percentage of coverage is rising with text database size. However, the graphs also flatten with increasing database size. In the Gutenberg Corpus, all graphs except the one for clustered diphones even decline when the text database is larger than 200,000 sentences. In contrast, in the plot for the Wikipedia corpus, the graphs are still rising at a text database size of 897,096 sentences. This indicates that using a larger part of the Wikipedia corpus may lead to an even higher coverage.

Generally, it can be observed that the coverage is higher when the Wikipedia corpus is used as selection basis.

Furthermore, the coverage of prosodic variations is considerably lower than the coverage of diphones. It also does not rise at the same extent as the diphone coverage. Obviously, there are significantly more sentences needed to raise the prosodic coverage than to raise the diphone coverage.

In Figure 7.2, the number of sentences that are needed to achieve full coverage for each of the four coverage measures is illustrated. The x-axis again

denotes the size of the text database and the y-axis denotes the size of the speech corpus after the stop criterion is reached. Each of the graphs in this

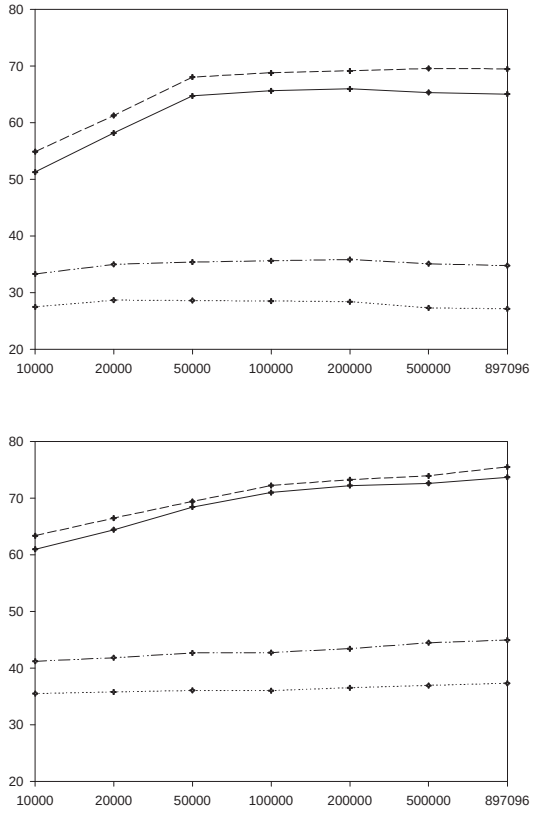

Figure 7.1: Coverage after 1000 sentences for sub corpora of different sizes of Gutenberg (top) and Wikipedia (bottom) corpus. Solid lines denote simple diphone coverage, dashed lines clustered diphone coverage, dotted lines simple prosody coverage and the lines with dots and dashes clustered prosody coverage.

figure corresponds to one of the four stop criteria simple diphones, simple prosody, clustered diphones and clustered prosody.

The number of sentences needed increases with text database size, but irregularly. As could be expected, the most sentences are needed when the simple prosody stop criterion is used. This is followed by the simple diphone stop criterion. The two graphs for simple/clustered diphone and simple/clustered prosody stop criterion, respectively, take a similar course, especially in the Wikipedia plot. Generally, the graphs for the prosody stop criteria are steeper.

The graphs show that for the Gutenberg corpus there are more sentences

needed to reach full coverage. This is consistent with the observation made above that the coverage is lower when the 1000 sentences are selected from sub corpora of the Gutenberg corpus and not from sub corpora of the Wikipedia corpus. The reason for this could be that the sentences in Wikipedia are longer on average (118.89 units in Wikipedia897096 versus 88.35 units in Gutenberg). Longer sentences are bound to contain more different units; thus more combinations can be covered with a long sentence than with a short sentence.

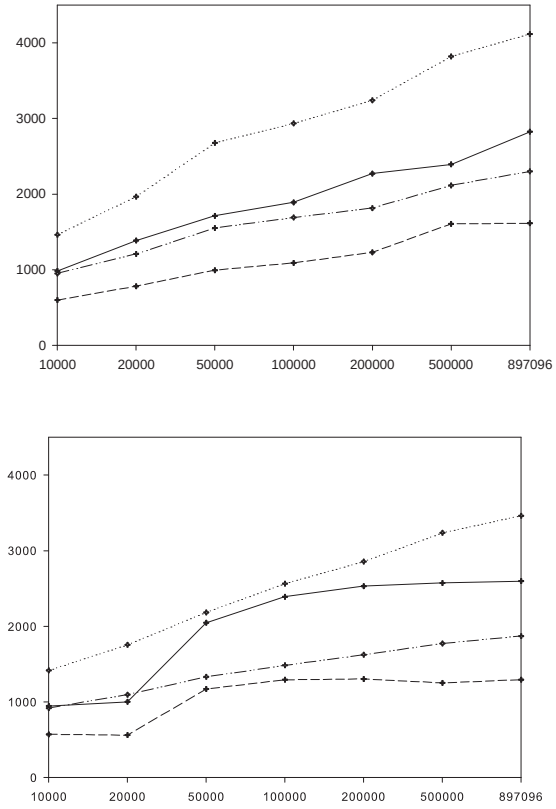

Figure 7.2: Number of sentences needed to fulfill the different stop criteria for sub corpora of different sizes of Gutenberg (top) and Wikipedia (bottom) corpus. Solid lines denote simple diphone stop criterion, dashed lines clustered diphone stop criterion, dotted lines simple prosody stop criterion and the lines with dots and dashes clustered prosody stop criterion.

As shown in the figure, even for a text corpus of only 10,000 sentences, about 1,500 sentences are needed to reach full simple prosody coverage. For

a text corpus of 897,096 sentences, 3,463 (Wikipedia) and 4,118 (Gutenberg) sentences are needed. Depending on the available resources, this is quite a large number for a synthesis voice.

Especially the graphs of simple and clustered diphone coverage in the Wikipedia plot are interesting: from 20,000 to 50,000 sentences the coverages escalate, and then the increase flattens. From a first glance one could conclude that in the 30,000 sentences that are added to the database, there are a lot of new diphones that have to be covered. But then the graphs for the prosody coverage would show the same characteristics, since new diphones also mean new prosodic variations.

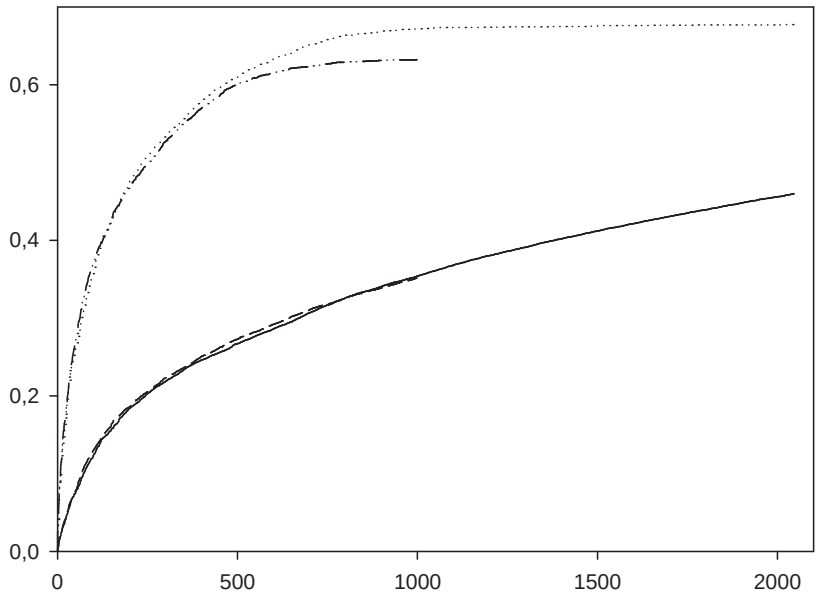

Figure 7.3: Coverage development for the text corpora selected from the first 20,000 and 50,000 sentences of the Wikipedia corpus with stop criterion simple diphones. The dashed line and the solid line represent simple prosody coverage development for 20,000 and 50,000 sentences, respectively. The line with dots and dashes and the dotted line denotes simple diphone coverage development for 20,000 and 50,000 sentences.

The reason for the increase becomes clearer when looking at the development of coverage in the cover set. Figure 7.3 shows the coverage development at the corresponding passes of the algorithm. The x-axis denotes the iterations of the

algorithm, while the y-axis denotes the coverage reached.

After a steep rise in the first 500 iterations, the graphs for the prosody coverage flatten slowly, while the graphs for the diphone coverage flatten out extremely, until a state is reached where the increase is almost zero. This stage is reached comparatively late for 20,000 sentences, while for 50,000 sentences, the graph shows a near zero growth for 1000 iterations.

The cause for this is reflected in the graphs for the prosody coverage: During the phase of the algorithm where the diphone coverage graph is nearly flat, the prosody graph still shows significant increase. So, obviously, while there are no new diphones added, new prosodic variations are added. This means that at this point in the algorithm, sentences with new prosodic variations get a higher score than sentences with new diphones.

A possible explanation for this is that, at the point in the algorithm where only a few diphones are still missing, it is unlikely that there is more than one new diphone in a sentence. At the same time, there are still a lot of new prosodic variations, and, because of their number, it is likely that a sentence contains more than one of them. Thus, the sentence with several new prosodic variations gets a higher score than the sentence with one new diphone.

7.3 Creation of example synthesis scripts

For the creation of an example script, the Wikipedia corpus was used. Section 7.3.1 describes the settings of the algorithm. In section 7.3.2, the resulting synthesis script is described. Unfortunately, the script falls short of the expectations, because it contains too many sentences that are unpronounceable for a German speaker. The reasons for this are discussed in section 7.3.3. As a result, the text corpus is reduced radically to increase sentence quality. The algorithm is then applied to the reduced corpus. The results of this pass are discussed in section 7.3.4.

7.3.1 Setup

The example script was produced by applying the algorithm to the first 897,096 sentences of the Wikipedia corpus. The parameter settings were the same as in the previous tests: units were simple diphones, wanted weights were set to 25/5/1 and divided by 1000, and the frequency weight was set to inverse. For this test, selection was stoped after 300 sentences are selected. Also, the sentence length was restricted: only sentences with more than 10 or less than 120 units were accepted.

7.3.2 Results of the first pass

The first pass of the algorithm produced disappointing results in terms of sentence quality. A lot of the selected sentences can not be used for recording, because they can not be pronounced by a German speaker.

Many of the sentences contained foreign words or foreign proper names. Most striking were Chinese characters and their transcription into Roman letters(Pinyin), but a number of other languages is also represented.

Examples for this are *Bis heute aktiv ist die 1903 gegründete Brauerei in Qīngdaǒ.* (*Until today, the brewery in Qīngdaǒ, founded in 1903, is active.*) and *Der Begriff Kohu rongorongo wird allgemein mit sprechendes Holz übersetzt.* (*The term Kohu rongorongo is generally translated with talking wood*).

Furthermore, a lot of unusual German terms can be found in the sentences. These include onomatopoeias: *Der häufigste Ruf ist ein unverkennbares, rollendes zizigürrrr oder gürrrr.* (*The most common cry is an unmistakable, rolling zizigürrr or gürrrr.*), and technical terms: *Diese Methodik hat, so Arendt, eine Nähe zur Sokratischen Mäeutik.* (*According to Arendt, this methodology is adjacent to the Socratic maieutics.*).

Moreover, some of the sentences are not real sentences. Either, some parts are missing, like in *Der Begriff Maui existiert als* (*The term Maui exists as*), or they contain Wikipedia format commands: *Er ist sechzehn Meter hoch. thumb /left /Auttiköngäs* (*It is 16 meters high. thumb /left /Auttiköngäs*).

These results call for the revision of the text corpus. This is done in the next section. The coverage of this first script is discussed in section 7.3.4.

7.3.3 Error analysis and resolution

It is obvious from the sentences listed in the previous section, that the Wikipedia corpus contains a number of sentences that should never have passed by the sentence credibility check described in section 5.5. The logical conclusion is that the credibility criteria used during the creation of the corpus were not strict enough. A higher threshold or lower values for unsafe transcription methods would have been better.

But this is only one aspect of the problem. Another aspect is an error in the categorization of the transcription methods. Apparently, the method "nothing" has to be subdivided, since the transcription is not only missing for punctuation, but also for words that can not be transcribed at all - like words in other alphabets. These words get a score of 100, which results in the corresponding sentence being rated much more credible than it deserves.

The only solution to these problems is to compute the sentence credibility anew for the whole 2,159,445 sentences of the Wikipedia corpus. This time, to

	Wikipedia897096	reduced
Number of sentences	897096	413401
Average sentence length	118.89 units	63.60 units
Maximum sentence length	10766 units	572 units
Minimum sentence length	1 unit	1 unit
phone coverage	100,00%	98.15%
simple diphone coverage	78.75%	58.52%
simple prosody coverage	60.05%	40.11%
clustered diphone coverage	77.92%	60.01%
clustered prosody coverage	62.63%	44.96%

Table 7.2: Statistics of first 897,096 sentences of the Wikipedia corpus and of the reduced Wikipedia corpus

be on the safe side, only words in the lexicon and punctuations are deemed as reliable.

The implementation was simplified: Instead of computing a score for each sentence, the sentence was rejected instantly, when a transcription method other than "lexicon" or "userdict" was used.

This method, of course, greatly reduces the size of the Wikipedia corpus: the resulting corpus contains only 413,401 sentences. Table 7.2 shows the statistics of the reduced corpus in comparison to the first 897,096 sentences of the full corpus. The table shows that the reduction in corpus size brings about a significant decrease in coverage: Diphone and prosody coverage is about 17-20% lower than in the full corpus. Phone coverage also decreases slightly.

The algorithm was applied to the reduced text corpus with the same settings as in the first try. The results are discussed in the next section.

7.3.4 Results of the second pass

The sentences of the speech corpus selected from the reduced text corpus are listed in appendix D.

The results show that the strict removal of unreliable sentences was worthwhile: All the problematic sentences described in section 7.3.2 have disappeared. Only the problem of incomplete sentences remains, rendering several sentences useless.

However, the number of sentences useful for recording has increased significantly: while for the first synthesis script, about 200 of the 300 sentences were useless, in the second synthesis script about 100 of the 300 sentences are

	pass 1	pass 2
Number of sentences	300	300
Average sentence length	66.64 units	43.01 units
Maximum sentence length	120 units	120 units
Minimum sentence length	10 units	10 units
phone coverage	100.00%	98.15%
simple diphone coverage	51.95%	43.84%
simple prosody coverage	20.92%	16.37%
clustered diphone coverage	58.78%	49.38%
clustered prosody coverage	27.94%	22.49%

Table 7.3: Coverage of selected sentences of first and second pass of the algorithm

useless.

So the quality of the sentences is all right now, but has the coverage suffered? Table 7.3 shows the distribution of the two passes of the algorithm in comparison.

In general, the coverage is very low for both sets. The reason for this is that too few sentences are selected.

Not surprisingly, the coverage is worse for the second pass than for the first. But the difference of the coverages of the two sets is not as large as the coverage differences of the two text databases from which the sentences are selected: While the prosody and diphone coverages are about 17-20% lower for the reduced text database than for the first 897,096 sentences of the Wikipedia corpus, the difference between the two scripts is between 4 and 9%.

The results show that distribution is not everything for a good speech corpus. The readability of the sentences has to be considered as well. A balance between the two extremes presented here - a huge corpus with good coverage and a lot of useless sentences, and a small corpus with worse coverage but mostly useful sentences - has to be found in order to produce a good speech corpus.

7.4 Summary

In this chapter, extensive tests were performed with the algorithm on the two German text corpora.

The goal of the tests desribed in section 7.1 is the search for the best settings of the algorithm. The results indicate that the inverse frequency measure leads

to the best distributions. As for the wanted weight settings, a preference for the diphone level gives good results for all coverage measures. For the prosody coverage measures, the setting of the wanted weight is not as important as the setting of the frequency weight. Setting the wanted weight to prefer new phones does not have much influence on the results.

In section 7.2, the amount of sentences needed to reach the maximum coverage for each of the four measures with different database sizes was examined. The results show that the number of sentences needed rises with database size. They also show that this effect diminishes with increasing text database size. This indicates that there is a maximum size for a text database, after which it is not useful to add new sentences anymore.

This is supported by the other tests conducted in this section, In these tests, the coverage reached when 1000 sentences are selected from databases of different sizes was measured. As in the previous tests, the coverage increases with corpus size, but the rise gets very flat in the end.

In section 7.3, an example synthesis script of 300 sentences was built using the algorithm. The first try, using the first 897,096 sentences of the Wikipedia corpus as the basis for selection, led to disappointing results in terms of sentence quality. The analysis of the selected sentences led to the conclusion that the sentence credibility check has to be stricter.

In consequence, a reduced Wikipedia corpus was built, with only those sentences that consisted entirely of words in the lexicon. Building a synthesis script from this text corpus yielded a synthesis script of much higher quality. However, the coverage of this script is worse than the coverage of the script with bad sentence quality.

8 Selection tools

The implementation of the algorithm used in this thesis is available as part of the Mary System. It comes with several other programs supporting the creation of a speech corpus. The programs are written in Java or Perl and can be started from the command line.

Apart from the selection program, which is described in detail in section 8.1, there is a program to build the text database, the basis for selection. It is described in section 8.2. Furthermore, section 8.3 describes a program for analyzing the results of the selection program. Finally, there are two Perl programs supporting the manual modification of a synthesis script. They are specified in section 8.4.

8.1 Selection Program

The selection program selects sentences from a text database according to the algorithm specified in this thesis. It needs three basic arguments: list of filenames, feature definition and stop criterion. The filenames constitute the text corpus in the form of binary feature files, whereas the feature definition supplies the basic information to read those feature files. The stop criterion is variable: the algorithm can stop after a specified number of sentences is selected, or after the maximum coverage is reached for one or more of the four different coverage measures simple diphones, simple prosody, clustered diphones and clustered prosody.

When the program is started for the first time, it reads in all sentences, builds the cover sets to be filled by the algorithm and computes the coverage statistics of the text corpus. A file containing all data needed for initialization of the program is stored on hard disk for subsequent runs. The text corpus statistics are also stored on disk to make the information available to the user.

The settings for the different parameters of the algorithm are read from a configuration file during the initialization of the program.

After the program has been initialized, the sentence selection is started. During selection, the progress of the program is printed to a log file and, if in verbose mode, also to the command line.

When the selection is completed, several files are written to disk. Firstly,

a file containing the resulting speech corpus in the form of a list of selected feature files. From this file, the synthesis script, that is, the list of the actual sentences of the corpus, can be created with the program presented in section 8.4. Secondly, the program creates a file containing the settings of the pass and the coverage of the speech corpus - for all four coverage measures and the phone coverage.

Apart from these basic functions, the selection program offers two options to influence the selection process: it can be given a set of sentences to include in the cover and another set of sentences to exclude from selection. This is useful, for example, for editing a synthesis script. Section 8.4 gives more details on this.

Another option of the selection program is to collect the settings and results of each pass of the algorithm in one log file. This log file can be analyzed by the analysis program described in section 8.3.

Finally, there is the option to log the development of coverage during the selection process. The result is a file containing a table with the phone, diphone and prosody coverage at every iteration of the algorithm. It can be used to construct a graph showing the coverage development over time, such as the graph in figure 7.3

8.2 Text database build program

The text database build program constructs the text database for the selection program out of a set of text files. The program heavily depends on Mary: For every file, the program uses Mary's text processing modules for splitting the given text into sentences. Then every sentence is checked for credibility, based on the phonetic transcription modules of Mary as described in section 7.3.3: All sentences containing words that were not transcribed with the use of the lexicon, the user lexicon or the preprocessing module are classified as unreliable. Additionally, there is the option to relax this rule a bit by rating also the words transcribed with the Denglish or the compound module as credible.

Unreliable sentences are written to a log file, whereas for the credible sentences, Mary is used again to compute the features for selection. Features and appropriate sentence are then stored on hard disk. The program collects the filenames of all credible feature files in one file. This list of sentences constitutes the text database and can be used by the selection program.

Due to format errors in the text, exceptions can occur during the execution of the program, causing it to abort. In order to avoid having to start the processing from the first file in the list again, the program keeps track of which files have already been processed. When it is restarted after a crash, it

resumes processing with the next file in the list.

8.3 Analysis program

The analysis program sorts the results of different passes of the program. The input of the program consists of one file containing the results of all passes. This file can be created with the selection program. The program produces six different files: One for each of the four coverage measures, one listing the passes that led to the same results and one sorting the results according to the number of selected sentences. In the files for the coverage measures, the results are sorted according to the coverage achieved for the current measure.

There are several modes of display for the results. In the full mode, settings and results are printed to the output files in every detail. In the justSettings mode, only the settings are printed, which makes the files more readable.

The program is useful to compare the performance of different settings of the algorithm. It was used for comparing the results of the tests in chapters 4 and 7, and for producing the tables in appendix C.

8.4 Script programs

There are two script programs: The first one produces a synthesis script from a set of selected feature files. The second one converts the synthesis script into two files: one lists the feature files of the useful sentences and the other one the feature files of the useless sentences. Whether a sentence is useful or useless, is determined by the user.

These programs are useful for manually correcting a synthesis script. The work flow is as follows: From the list of feature files selected by the selection program, a synthesis script is created with the first program. The user then looks over the sentences and sorts out those that are not suitable for recording. Examples for unsuitable sentences are sentences that are not complete, contain ambiguous words or words that are difficult to pronounce. Using the second script program, the script is divided into lists of good and bad feature files. Now the selection program is called with the good feature files to be included in the speech corpus and the bad feature files to be removed from the text database prior to selection. The program creates a new list of features which can be converted to a synthesis script. This process can be repeated until the synthesis script is completed.

9 Summary and Conclusion

In this thesis, an algorithm for selecting sentences for a speech corpus has been proposed, implemented and tested thoroughly. The algorithm is available as part of the Mary Text-To-Speech System. The system can be downloaded from `http://mary.dfki.de` and is released as open source.

The algorithm is first described in chapter 3. It is a variant of the greedy algorithm: At every iteration of the algorithm, the sentence with the highest sentence score is selected. In the first version of the algorithm, there was also a variant in which the sentence with the highest score and the highest number of new units (normalized by sentence length) was selected. This variant was dropped in the final implementation, because it proved to be too unreliable in the first tests.

For the computation of sentence score, the score of the units in a sentence is summed up and divided by sentence length. There are two different definitions of a unit: simple diphone and clustered diphone. Simple diphone units consist of the three features "phone", "next phone" and "prosody". Analogous to that, clustered diphones contain the features "phone", "next phone class" and "prosody". A phone class is a set of one or more phones that have a similar place of articulation, like, for example, alveolar consonants. Phone classes are used in order to reduce the number of diphone combinations.

Unit score is the sum of the scores of the features. For every feature value, the score is the product of frequency weight and wanted weight. The frequency weight reflects the frequency of the current value in the text database. There are four different frequency measures: relative frequency, 1 minus relative frequency (1minus), 1 divided by relative frequency (inverse frequency), and 1 (no frequency). The wanted weight reflects how much a value is wanted in the speech corpus. It is decreased every time a unit with that value is added to the speech corpus.

As described in chapter 4, the algorithm was first tested on the database of the Blizzard Challenge, an English database consisting of text and appropriate recordings. Based on the results of the tests, the algorithm was optimized as described in chapter 5. At the same time, from the results of the tests, the set of sentences with the best distribution was chosen to build a voice, which was then submitted to the Blizzard Challenge. In the challenge, the mean opinion score of this voice was worse than the voice built from the full database, but

better than the voice built from the Arctic subset of the database.

Apart from the optimization of the algorithm, chapter 5 also describes the preparations made to cope with the German text corpora that are described in chapter 6. Most notably, a sentence credibility check is introduced to rule out sentences whose phonetic transcriptions are questionable.

Chapter 6 describes the building process of two German text corpora. They were built from Internet resources: one of the corpora, called the Gutenberg corpus, consists of the German books of Project Gutenberg (Gutenberg (2007)), the other one, Wikipedia corpus, consists of articles from the German Wikipedia (Wikipedia (2007a)).

Both text corpora were first downloaded and stored in text files. Then these files were divided into sentences with the text processing modules of Mary. For the Wikipedia corpus, major removal of formatting commands had to be performed. Unreliable sentences were removed with the above mentioned sentence credibility check and the unit features were computed with the help of Mary. The resulting corpora consist of 897,096 sentences (Gutenberg) and 2,159,445 sentences (Wikipedia), respectively.

The corpora were used as a basis for further tests on the algorithm, described in chapter 7. First of all, the best settings for the algorithm were investigated by applying the algorithm to the first 5000 sentences of each corpus with different settings. At each pass, 500 sentences were selected and their distribution measured.

The results indicate that, for the wanted weight, weighting the features "next phone" and "next phone class", respectively, 5 times higher than the prosody feature is a good setting for achieving good diphone coverage. For prosodic coverage, there seems to be no ultimate best setting for the wanted weight. However, the performance of the algorithm is most dependent on the frequency weight, with inverse frequency being the best setting, and relative frequency the worst.

The second round of tests centered on the topic of corpus size: How large does the text corpus have to be to ensure a good coverage of the speech corpus? For answering this question, the tests were conducted on sub corpora of different sizes of the Gutenberg and the Wikipedia corpus. The coverage of 1000 selected sentences was investigated, as well as the number of sentences needed for achieving full coverage.

The results show that, although the achievable coverage of the speech corpus rises with text corpus size, after the text corpus has exceeded a certain size, the benefit of adding more sentences to the text corpus decreases. Also, the number of sentences needed to achieve full coverage does not necessarily reflect the real number of new unit types.

Finally, chapter 7 also describes the attempt to build a synthesis script (the actual list of sentences of a speech corpus) from the Wikipedia corpus. The first attempt failed, because there were too many unreliable sentences in the resulting synthesis script. The sentence credibility check used during the creation of the corpus obviously was too lax. Therefore, the sentences of the Wikipedia corpus are checked again more strictly, reducing the Wikipedia corpus from 2,159,445 to 413,401 sentences. The synthesis script selected from the new text corpus is significantly more suitable for recording. The coverage of the reduced text corpus is about 17-20% lower than the coverage of the Wikipedia corpus, but the coverage of the synthesis script gained from this corpus is only 4-9% lower than the coverage of the script obtained from the full corpus.

Chapter 8 describes the implementation of the algorithm in the form that it is available as part of Mary. In addition to the actual selection program, several other programs are available as well. These programs can be used to create a text database or to handle the results of the selection. Thus, they comprise a whole selection toolkit.

The question if the algorithm presented in this thesis is capable of selecting the optimal speech corpus is not easy to answer. First of all, the meaning of the term "optimal" in this context has to be further specified. The optimal speech corpus has the maximum coverage for all four coverage measures. At the same time, the number of sentences in the corpus is the minimum number of sentences needed to reach the maximum coverage. Furthermore, the quality of all sentences is high enough for them to be recorded.

Apart from the coverage of the speech corpus, those conditions are not easy to assess. It is not clear what the minimum number of sentences needed is. The experiments with different text corpora sizes in section 7.2 show that the coverage attainable and the number of sentences needed is different for different text corpora.

The quality of the sentences in the speech corpus is very dependent on the text corpus. As the problematic sentences that occur in the example synthesis script described in section 7.3 show, preselection of the sentences in the text corpus is very important. In addition to a criterion for sentence credibility, the texts from which the corpus is built should be chosen with care. This way, undesirable sentences like those removed by the sentence credibility check in section 6.2 can be ruled out from the beginning.

So, for the evaluation of the algorithm, the coverage measures are the most important factor. The tests in chapter 7 show that the maximum coverage can be achieved with the algorithm. In particular, the tests in section 7.1 also show that the performance of the algorithm depends heavily on the settings

of the parameters. There has to be a balance between frequency weight and wanted weight to reach a good coverage.

However, section 7.2 shows that the selection is not always optimal. For some passes of the algorithm, more sentences than needed for maximum diphone coverage were selected, presumably because new prosodic variations were rated higher than new diphones.

Therefore, it might make sense to reduce the speech corpus produced by the algorithm. This could be done, for example, with the inverse greedy algorithm that is described in François and Boëffard (2002).

All in all, it was shown that the presented algorithm is suitable for selecting a speech corpus. With the selection tools created in this thesis, there now exists a basis to create a corpus for a synthesis voice with adequate coverage.

Bibliography

Ove Andersen and Charles Hoequist. Keeping Rare Events Rare. In *Proceedings Eurospeech, Geneva, Switzerland*, pages 1337–1340, 2003.

Alan W. Black and Kevin A. Lenzo. Optimal Data Selection for Unit Selection Synthesis. In *4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland*, pages 63–67, 2001.

Baris Bozkurt, Ozlem Ozturk, and Thierry Dutoit. Text Design for TTS Speech Corpus Building using a Modified Greedy Selection. In *Proceedings Eurospeech, Geneva, Switzerland*, pages 277–280, 2003.

Hélène François and Olivier Boëffard. The Greedy Algorithm and its Application to the Construction of a Continuous Speech Database. In *Proceedings LREC, Las Palmas, Canary Islands, Spain*, 2002.

Jean-Luc Gauvain, Lori Lamel, and Maxine Eskénazi. Design Considerations and Text Selection for BREF, a large French Read Corpus. In *Proceedings ICSLP, Kobe, Japan,*, pages 1097–1100, 1990.

Project Gutenberg. German Ebooks of Project Gutenberg, 2007. URL `http://www.gutenberg.org/browse/languages/de`. Online; as of 17 th of April, 2007.

John Kominek and Alan Black. Cmu Arctic databases for Speech Synthesis. Technical report, CMU Language Technologies Institute, 2003.

Bernd Möbius. Rare Events and Closed Domains - Two Delicate Concepts in Speech Synthesis. *International Journal of Speech Technology*, 6:57–71, 2003.

Marc Schröder and Jürgen Trouvain. The German Text-To-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6:365–377, 2003.

Antje Schweitzer, Norbet Braunschweiler, Tanja Klankert, Bernd Möbius, and Bettina Sauberlich. Restricted Unlimited Domain Synthesis. In *Proceedings Eurospeech, Geneva, Switzerland*, pages 1321–1324, 2003.

Chilin Shih and Benjamin Ao. Duration Study for the AT&T Mandarin Text-To-Speech System. In *Second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, NY, USA*, pages 29–32, 1994.

Jan P. H. van Santen and Adam L. Buchsbaum. Methods for Optimal Text Selection. In *Proceedings Eurospeech, Rhodes, Greece*, pages 553–556, 1997.

Jan P.H. van Santen. Combinatorial Issues in Text-To-Speech Synthesis. In *Proceedings Eurospeech, Rhodes, Greece*, pages 2507–2510, 1997.

Wikipedia. Data Dump of German Wikipedia, 2007a. URL `http://dumps.wikimedia.org/backup-index.html`. Online; dump from 27 th of April, 2007.

Wikipedia. Levenshtein Distance — Wikipedia, The Free Encyclopedia, 2007b. URL `http://en.wikipedia.org/w/index.php?title=Levenshtein_distance&oldid=141557929`. Online; version from 30th of June, 2007.

Wikipedia. Mean Opinion Score — Wikipedia, The Free Encyclopedia, 2007c. URL `http://en.wikipedia.org/w/index.php?title=Mean_Opinion_Score&oldid=141369794`. Online; version from 29th of June, 2007.

Wikipedia. Word Error Rate — Wikipedia, The Free Encyclopedia, 2006. URL `http://en.wikipedia.org/w/index.php?title=Word_error_rate&oldid=85646701`. Online; version from 4th of November, 2006.

A English phone classes

class	Sampa	IPA
0	0,_	0,pause
c_labial	b,m,p,w	b,m,p,w
c_alveolar	d,l,n,r,s,t,z	d,l,n,r,s,t,z
c_palatal	tS,DZ,S,j,Z	tʃ,ðʒ,ʃ,j,ʒ
c_labiodental	f,v	f,v
c_dental	D,T	ð,θ
c_velar	g,k,N	g,k,ŋ
c_glottal	h	h
v_i	I,i	ɪ,i
v_u	U,u	ʊ,u
v_o	A,O	ɑ,ɔ
v_E	E	ɛ
v_EI	EI	ɛɪ
v_V	V	ʌ
v_@	@	ə
v_r=	r=	r=
v_@U	@U	əʊ
v_OI	OI	ɔɪ
v_{	{	ɐ
v_aU	aU	aʊ
v_AI	AI	ɑɪ

Table A.1: The 21 phone classes for English

B German phone classes

class	Sampa	IPA
0	0,_,?	0,pause.P
c_labial	b,m,p,w,pf	b,m,p,w,pf
c_alveolar	d,l,n,r,s,t,z	d,l,n,r,s,t,z
c_palatal	tS,S,j,Z	tʃ,ʃ,j,ʒ
c_labiodental	f,v	f,v
c_dental	D,T	ð,θ
c_velar	g,k,N,C	g,k,ŋ,ɕ
c_glottal	h	h
c_uvular	x,K	x,ʁ
v_i	i,i:	i,iː
v_u	U,u,u:	ʊ,u,uː
v_O	O	ɔ
v_o	o,o:	o,oː
v_E	E,E:	ɛ,ɛː
v_EI	EI	ɛɪ
v_@	@	ə
v_aU	AU	aʊ
v_6	6	ɒ
v_~	a~,e~,o~,9~	ã,ẽ,õ,œ̃
v_a	a,a:	a,aː
v_y	y,Y	y,ʏ
v_2	2,2:	ø,øː
v_e	e,e:	e,eː
v_9	9	œ
v_OY	OY	ɔʏ
v_Ya	Ya	ʏa
v_aI	aI	aɪ

Table B.1: The 27 phone classes for German

C Results of the setting tests

The following tables show the results of the tests for the best settings. Each setting was tested with the first 5000 sentences of the Gutenberg corpus and the Wikipedia corpus, respectively. The ordering of the tables reflects the results for the different settings. The best settings are at the top.

In table C.1, the settings are ordered in respect to with which setting the highest simple diphone coverage was achieved. Analogously, table C.2 sorts the settings according to clustered diphone coverage achieved. Tables C.3 and C.4 do the same for simple prosody coverage and clustered prosody coverage, respectively.

Abbrevations:

- SD/CD = units are simple/clustered diphones
- inverse;minus1;none;normal = use inverse, minus1, none or normal frequency weight
- 25/5/1 = wanted weights: phone level 25, diphone level 5, prosody level 1
- 2;5;1000 = divide wanted weight by 2, 5 or 1000

Best settings for simple diphone coverage					
Gutenberg			Wikipedia		
	Settings	Coverage max 0.43670		Settings	Coverage max 0.56830
1	SD;inverse;25/5/1;5	0.42997	1	SD;inverse;25/5/1;1000	0.55791
1	SD;inverse;25/5/1;1000	0.42997	1	SD;inverse;1/5/1;1000	0.55791
1	SD;inverse;1/5/1;5	0.42997	2	SD;inverse;25/5/1;5	0.5569
1	SD;inverse;1/5/1;1000	0.42997	2	SD;inverse;1/5/1;5	0.5569
2	SD;inverse;25/5/1;2	0.42727	3	SD;inverse;25/5/1;2	0.55488
2	SD;inverse;1/5/1;2	0.42727	3	SD;inverse;1/5/1;2	0.55488
3	SD;1minus;25/5/1;1000	0.41515	4	SD;none;1/5/1;1000	0.54276
4	SD;none;1/5/1;1000	0.41347	4	SD;1minus;1/5/1;1000	0.54276
5	SD;none;25/5/1;1000	0.41313	5	SD;1minus;25/5/1;1000	0.54141
6	SD;none;25/5/1;5	0.41279	6	SD;none;25/5/1;1000	0.54108
6	SD;none;1/5/1;5	0.41279	7	SD;none;25/5/1;5	0.53939
6	SD;1minus;1/5/1;1000	0.41279	8	SD;1minus;1/5/1;5	0.53805
7	SD;1minus;25/5/1;5	0.41111	9	SD;none;1/5/1;5	0.53737

	Settings	Coverage max 0.43670		Settings	Coverage max 0.56830
8	SD;1minus;1/5/1;5	0.4101	9	SD;1minus;25/5/1;5	0.53737
9	SD;inverse;1/1/1;5	0.40875	10	SD;inverse;1/1/1;1000	0.53502
10	SD;inverse;1/1/1;1000	0.40842	11	SD;inverse;1/1/1;5	0.53401
11	SD;inverse;1/1/1;2	0.40741	12	SD;inverse;1/1/1;2	0.53098
12	SD;1minus;25/5/1;2	0.40135	13	SD;1minus;25/5/1;2	0.52626
12	SD;1minus;1/5/1;2	0.40135	14	SD;1minus;1/5/1;2	0.52525
13	SD;none;25/5/1;2	0.40067	15	SD;none;25/5/1;2	0.52424
13	SD;none;1/5/1;2	0.40067	16	CD;inverse;25/5/1;2	0.5229
14	CD;inverse;25/5/1;2	0.39596	16	CD;inverse;1/5/1;2	0.5229
14	CD;inverse;1/5/1;2	0.39596	17	SD;none;1/5/1;2	0.52222
15	SD;inverse;1/5/25;5	0.39259	18	SD;inverse;1/5/25;1000	0.5202
16	SD;inverse;1/5/25;1000	0.39226	19	CD;inverse;25/5/1;5	0.51953
17	SD;inverse;1/5/25;2	0.39192	19	CD;inverse;1/5/1;5	0.51953
17	CD;inverse;25/5/1;5	0.39192	20	SD;inverse;1/5/25;5	0.51852
17	CD;inverse;1/5/1;5	0.39192	21	SD;inverse;1/5/25;2	0.51717
18	CD;inverse;25/5/1;1000	0.39057	21	CD;none;25/5/1;2	0.51717
18	CD;inverse;1/5/1;1000	0.39057	21	CD;none;1/5/1;2	0.51717
18	CD;inverse;1/1/1;5	0.39057	22	CD;1minus;25/5/1;2	0.51684
19	CD;inverse;1/1/1;2	0.3899	23	CD;inverse;1/1/1;2	0.5165
19	CD;inverse;1/1/1;1000	0.3899	24	CD;1minus;25/5/1;5	0.51616
20	CD;inverse;1/5/25;5	0.38687	24	CD;1minus;1/5/1;2	0.51616
21	CD;inverse;1/5/25;2	0.38586	25	CD;1minus;1/5/1;5	0.51582
22	CD;inverse;1/5/25;1000	0.38519	26	CD;inverse;25/5/1;1000	0.51549
23	SD;1minus;1/1/1;5	0.38215	26	CD;inverse;1/5/1;1000	0.51549
24	CD;none;1/5/1;2	0.38047	27	CD;none;25/5/1;5	0.51515
25	SD;none;1/1/1;5	0.3798	27	CD;none;1/5/1;5	0.51515
25	SD;1minus;1/1/1;1000	0.3798	28	CD;inverse;1/1/1;5	0.51279
26	SD;none;1/1/1;2	0.37912	29	CD;1minus;1/5/1;1000	0.51246
26	SD;1minus;1/1/1;2	0.37912	30	CD;none;25/5/1;1000	0.51212
27	CD;none;25/5/1;5	0.37879	31	CD;inverse;1/1/1;1000	0.51178
27	CD;none;25/5/1;1000	0.37879	32	CD;none;1/5/1;1000	0.5101
27	CD;1minus;25/5/1;5	0.37879	32	CD;inverse;1/5/25;2	0.5101
27	CD;1minus;25/5/1;1000	0.37879	33	CD;1minus;25/5/1;1000	0.50842
28	CD;1minus;1/5/1;5	0.37845	34	CD;inverse;1/5/25;1000	0.50741
29	SD;none;1/1/1;1000	0.37811	35	SD;1minus;1/1/1;1000	0.50707
29	CD;none;1/5/1;5	0.37811	36	SD;1minus;1/1/1;5	0.50673
29	CD;none;1/5/1;1000	0.37811	37	CD;inverse;1/5/25;5	0.5064

	Settings	Coverage max 0.43670		Settings	Coverage max 0.56830
30	CD;1minus;1/5/1;2	0.37744	38	SD;none;1/1/1;5	0.50606
31	CD;1minus;1/5/1;1000	0.37677	39	SD;none;1/1/1;1000	0.50269
32	CD;none;25/5/1;2	0.37643	40	SD;none;1/1/1;2	0.49899
33	CD;1minus;25/5/1;2	0.37576	41	CD;none;1/1/1;5	0.49697
34	SD;1minus;1/5/25;2	0.36431	42	SD;1minus;1/1/1;2	0.4963
35	SD;1minus;1/5/25;1000	0.36364	43	CD;1minus;1/1/1;1000	0.49461
36	SD;none;1/5/25;2	0.36296	44	CD;1minus;1/1/1;5	0.49428
37	SD;none;1/5/25;5	0.36195	45	CD;none;1/1/1;1000	0.49394
37	CD;1minus;1/1/1;5	0.36195	46	SD;none;1/5/25;1000	0.48822
38	CD;none;1/1/1;2	0.36162	46	CD;1minus;1/1/1;2	0.48822
39	SD;1minus;1/5/25;5	0.36094	47	SD;none;1/5/25;5	0.48788
40	CD;none;1/1/1;5	0.36061	48	CD;none;1/1/1;2	0.48687
40	CD;1minus;1/1/1;2	0.36061	49	SD;1minus;1/5/25;1000	0.4862
41	SD;none;1/5/25;1000	0.36027	50	SD;1minus;1/5/25;5	0.48586
42	CD;none;1/1/1;1000	0.3596	51	CD;none;1/5/25;1000	0.48013
43	CD;1minus;1/1/1;1000	0.35791	52	CD;1minus;1/5/25;1000	0.4798
44	SD;normal;25/5/1;1000	0.35623	53	SD;1minus;1/5/25;2	0.47912
45	SD;normal;1/5/1;1000	0.35387	54	CD;none;1/5/25;2	0.47845
46	CD;none;1/5/25;5	0.35152	54	CD;1minus;1/5/25;5	0.47845
47	SD;normal;1/5/1;5	0.35118	55	SD;normal;25/5/1;1000	0.47811
48	CD;none;1/5/25;2	0.35017	56	CD;none;1/5/25;5	0.47778
49	CD;1minus;1/5/25;1000	0.34781	57	SD;none;1/5/25;2	0.47744
50	CD;1minus;1/5/25;2	0.34714	57	SD;normal;25/5/1;5	0.47744
51	CD;1minus;1/5/25;5	0.34613	58	SD;normal;1/5/1;5	0.47576
52	CD;none;1/5/25;1000	0.34444	59	SD;normal;1/5/1;1000	0.47542
53	SD;normal;25/5/1;5	0.34411	60	CD;1minus;1/5/25;2	0.47508
54	CD;normal;1/5/1;1000	0.34209	61	CD;normal;25/5/1;5	0.47441
55	CD;normal;25/5/1;1000	0.34141	62	CD;normal;25/5/1;1000	0.4734
56	CD;normal;25/5/1;5	0.33939	62	CD;normal;1/5/1;5	0.4734
57	CD;normal;1/5/1;2	0.3367	63	CD;normal;1/5/1;1000	0.47306
58	SD;normal;25/5/1;2	0.33502	64	CD;normal;25/5/1;2	0.46465
59	CD;normal;1/5/1;5	0.33468	65	SD;normal;25/5/1;2	0.46027
60	CD;normal;25/5/1;2	0.33434	66	CD;normal;1/1/1;1000	0.45993
61	SD;normal;1/5/1;2	0.33165	67	SD;normal;1/1/1;5	0.45758
62	SD;normal;1/1/1;1000	0.32929	67	SD;normal;1/1/1;1000	0.45758
63	SD;normal;1/1/1;5	0.3266	68	SD;normal;1/5/1;2	0.45724
64	CD;normal;1/1/1;5	0.32424	69	CD;normal;1/5/1;2	0.4569

	Settings	Coverage max 0.43670		Settings	Coverage max 0.56830
64	CD;normal;1/1/1;1000	0.32424	70	CD;normal;1/1/1;5	0.45084
65	SD;normal;1/5/25;5	0.31818	71	CD;normal;1/1/1;2	0.4468
66	SD;normal;1/5/25;1000	0.31785	72	SD;normal;1/1/1;2	0.44613
66	CD;normal;1/1/1;2	0.31785	73	CD;normal;1/5/25;1000	0.44512
67	CD;normal;1/5/25;5	0.3165	74	SD;normal;1/5/25;5	0.44108
68	CD;normal;1/5/25;1000	0.31515	75	CD;normal;1/5/25;5	0.44074
69	SD;normal;1/1/1;2	0.31448	76	SD;normal;1/5/25;1000	0.43939
70	CD;normal;1/5/25;2	0.30976	77	CD;normal;1/5/25;2	0.4367
71	SD;normal;1/5/25;2	0.30135	78	SD;normal;1/5/25;2	0.43064

Table C.1: Best settings for simple diphone coverage

Best settings for clustered diphone coverage					
Gutenberg			Wikipedia		
	Settings	Coverage max 0.47531		Settings	Coverage max 0.59945
1	CD;inverse;25/5/1;5	0.47531	1	CD;inverse;25/5/1;5	0.59945
1	CD;inverse;25/5/1;1000	0.47531	1	CD;inverse;25/5/1;1000	0.59945
1	CD;inverse;1/5/1;5	0.47531	1	CD;inverse;1/5/1;5	0.59945
1	CD;inverse;1/5/1;1000	0.47531	1	CD;inverse;1/5/1;1000	0.59945
2	CD;inverse;25/5/1;2	0.47394	2	CD;inverse;25/5/1;2	0.59877
2	CD;inverse;1/5/1;2	0.47394	2	CD;inverse;1/5/1;2	0.59877
3	CD;none;25/5/1;1000	0.47325	3	SD;inverse;25/5/1;5	0.59602
3	CD;1minus;25/5/1;1000	0.47325	3	SD;inverse;25/5/1;1000	0.59602
4	CD;none;1/5/1;1000	0.47257	3	SD;inverse;1/5/1;5	0.59602
5	CD;1minus;1/5/1;1000	0.47188	3	SD;inverse;1/5/1;1000	0.59602
6	CD;inverse;1/1/1;5	0.47119	4	SD;inverse;25/5/1;2	0.59534
6	CD;inverse;1/1/1;1000	0.47119	4	SD;inverse;1/5/1;2	0.59534
6	CD;1minus;25/5/1;5	0.47119	4	CD;1minus;1/5/1;1000	0.59534
7	SD;inverse;25/5/1;5	0.47051	5	CD;none;1/5/1;1000	0.59465
7	SD;inverse;1/5/1;5	0.47051	5	CD;1minus;25/5/1;5	0.59465
8	SD;inverse;25/5/1;1000	0.46982	6	CD;none;25/5/1;1000	0.59396
8	SD;inverse;1/5/1;1000	0.46982	6	CD;inverse;1/1/1;5	0.59396
8	CD;none;25/5/1;5	0.46982	7	CD;inverse;1/1/1;1000	0.59328
8	CD;none;1/5/1;5	0.46982	7	CD;1minus;25/5/1;1000	0.59328
8	CD;1minus;1/5/1;5	0.46982	7	CD;1minus;1/5/1;5	0.59328
9	CD;inverse;1/1/1;2	0.46914	8	CD;none;25/5/1;5	0.59259

	Settings	Coverage max 0.47531		Settings	Coverage max 0.59945
10	SD;inverse;25/5/1;2	0.46845	8	CD;none;1/5/1;5	0.59259
10	SD;inverse;1/5/1;2	0.46845	9	CD;inverse;1/1/1;2	0.59191
11	CD;inverse;1/5/25;5	0.46502	10	CD;none;25/5/1;2	0.58848
11	CD;inverse;1/5/25;1000	0.46502	11	CD;none;1/5/1;2	0.58779
12	CD;none;1/5/1;2	0.46365	12	CD;1minus;25/5/1;2	0.58711
13	CD;inverse;1/5/25;2	0.46296	12	CD;1minus;1/5/1;2	0.58711
13	CD;1minus;25/5/1;2	0.46296	13	CD;inverse;1/5/25;2	0.58368
14	CD;none;25/5/1;2	0.46228	13	CD;inverse;1/5/25;1000	0.58368
14	CD;1minus;1/5/1;2	0.46228	14	CD;inverse;1/5/25;5	0.58299
15	SD;none;1/5/1;1000	0.46022	15	SD;1minus;1/5/1;1000	0.57956
16	SD;none;25/5/1;1000	0.45885	16	SD;1minus;25/5/1;1000	0.57888
17	SD;1minus;1/5/1;1000	0.45816	17	SD;none;1/5/1;1000	0.57819
18	SD;none;1/5/1;5	0.45748	18	SD;none;25/5/1;1000	0.5775
19	SD;1minus;25/5/1;1000	0.45679	19	SD;none;25/5/1;5	0.57613
20	SD;none;25/5/1;5	0.4561	19	SD;inverse;1/1/1;1000	0.57613
21	SD;1minus;25/5/1;5	0.45542	20	SD;1minus;1/5/1;5	0.57545
22	SD;1minus;1/5/1;5	0.45405	21	SD;inverse;1/1/1;5	0.57476
23	SD;inverse;1/1/1;2	0.45199	22	SD;none;1/5/1;5	0.5727
23	SD;inverse;1/1/1;5	0.45199	22	SD;1minus;25/5/1;5	0.5727
24	SD;inverse;1/1/1;1000	0.4513	23	SD;inverse;1/1/1;2	0.57133
25	SD;none;25/5/1;2	0.4465	24	CD;none;1/1/1;1000	0.56653
25	SD;1minus;1/5/1;2	0.4465	24	CD;1minus;1/1/1;1000	0.56653
25	CD;1minus;1/1/1;5	0.4465	25	SD;1minus;25/5/1;2	0.56447
26	SD;1minus;25/5/1;2	0.44513	25	CD;none;1/1/1;5	0.56447
26	CD;none;1/1/1;1000	0.44513	26	CD;1minus;1/1/1;5	0.56379
27	SD;none;1/5/1;2	0.44444	27	SD;none;25/5/1;2	0.5631
27	CD;1minus;1/1/1;1000	0.44444	27	SD;1minus;1/5/1;2	0.5631
28	CD;none;1/1/1;5	0.44307	28	SD;inverse;1/5/25;1000	0.56173
29	SD;inverse;1/5/25;5	0.4417	29	SD;none;1/5/1;2	0.56036
30	SD;inverse;1/5/25;2	0.44102	29	SD;inverse;1/5/25;5	0.56036
30	SD;inverse;1/5/25;1000	0.44102	30	SD;inverse;1/5/25;2	0.55693
31	CD;none;1/1/1;2	0.43896	31	CD;none;1/1/1;2	0.5487
32	CD;1minus;1/1/1;2	0.4369	32	CD;1minus;1/1/1;2	0.54733
33	SD;1minus;1/1/1;5	0.43278	33	SD;none;1/1/1;5	0.54595
34	SD;none;1/1/1;5	0.4321	33	SD;1minus;1/1/1;5	0.54595
35	SD;1minus;1/1/1;1000	0.43073	34	SD;1minus;1/1/1;1000	0.54527
36	SD;none;1/1/1;1000	0.43004	35	CD;none;1/5/25;1000	0.54321

	Settings	Coverage max 0.47531		Settings	Coverage max 0.59945
37	CD;normal;25/5/1;1000	0.42867	36	CD;normal;25/5/1;1000	0.54115
38	CD;normal;1/5/1;1000	0.42798	36	CD;1minus;1/5/25;5	0.54115
39	CD;none;1/5/25;5	0.42661	36	CD;1minus;1/5/25;1000	0.54115
40	SD;1minus;1/1/1;2	0.42524	37	CD;normal;1/5/1;1000	0.54047
41	SD;none;1/1/1;2	0.42318	38	CD;none;1/5/25;5	0.53978
41	CD;1minus;1/5/25;1000	0.42318	39	SD;none;1/1/1;2	0.53909
42	CD;none;1/5/25;1000	0.42181	40	SD;none;1/1/1;1000	0.53841
43	CD;1minus;1/5/25;5	0.42112	41	CD;normal;1/5/1;5	0.53772
44	CD;1minus;1/5/25;2	0.42044	42	CD;normal;25/5/1;5	0.53567
45	CD;normal;25/5/1;5	0.41975	43	SD;1minus;1/1/1;2	0.53361
46	SD;1minus;1/5/25;1000	0.41838	44	CD;none;1/5/25;2	0.53155
46	CD;none;1/5/25;2	0.41838	45	CD;1minus;1/5/25;2	0.52881
47	SD;1minus;1/5/25;2	0.41495	46	SD;none;1/5/25;1000	0.52606
48	SD;none;1/5/25;5	0.41358	47	SD;none;1/5/25;5	0.52538
48	SD;none;1/5/25;1000	0.41358	48	SD;1minus;1/5/25;1000	0.52126
49	SD;1minus;1/5/25;5	0.41289	48	CD;normal;1/1/1;1000	0.52126
49	CD;normal;1/5/1;5	0.41289	49	SD;1minus;1/5/25;5	0.5192
50	SD;none;1/5/25;2	0.41152	50	SD;1minus;1/5/25;2	0.51852
51	SD;normal;25/5/1;1000	0.40741	51	SD;none;1/5/25;2	0.5144
52	SD;normal;1/5/1;1000	0.40604	51	CD;normal;25/5/1;2	0.5144
53	SD;normal;1/5/1;5	0.40329	52	SD;normal;25/5/1;5	0.51303
54	CD;normal;25/5/1;2	0.39918	52	SD;normal;25/5/1;1000	0.51303
54	CD;normal;1/1/1;1000	0.39918	53	SD;normal;1/5/1;5	0.51097
55	CD;normal;1/5/1;2	0.39849	53	SD;normal;1/5/1;1000	0.51097
56	SD;normal;25/5/1;5	0.39506	54	CD;normal;1/5/1;2	0.51029
57	CD;normal;1/1/1;5	0.393	55	CD;normal;1/1/1;5	0.50686
58	SD;normal;25/5/1;2	0.38889	56	SD;normal;1/1/1;5	0.4952
59	SD;normal;1/1/1;1000	0.38203	57	CD;normal;1/1/1;2	0.49451
60	SD;normal;1/5/1;2	0.38134	58	SD;normal;25/5/1;2	0.49383
61	CD;normal;1/5/25;1000	0.38066	58	CD;normal;1/5/25;1000	0.49383
62	SD;normal;1/1/1;5	0.37929	59	SD;normal;1/5/1;2	0.49108
63	CD;normal;1/5/25;5	0.37723	60	SD;normal;1/1/1;1000	0.48971
64	CD;normal;1/1/1;2	0.37311	61	CD;normal;1/5/25;5	0.4856
65	SD;normal;1/5/25;1000	0.37174	62	SD;normal;1/1/1;2	0.48148
66	SD;normal;1/5/25;5	0.37037	63	SD;normal;1/5/25;1000	0.48011
67	CD;normal;1/5/25;2	0.36968	64	CD;normal;1/5/25;2	0.47942
68	SD;normal;1/1/1;2	0.36557	65	SD;normal;1/5/25;5	0.47599

	Settings	Coverage max 0.47531		Settings	Coverage max 0.59945
69	SD;normal;1/5/25;2	0.35528	66	SD;normal;1/5/25;2	0.46571

Table C.2: Best settings for clustered diphone coverage

Best settings for simple prosody coverage					
Gutenberg			Wikipedia		
	Settings	Coverage max 0.25387		Settings	Coverage max 0.34815
1	SD;inverse;1/1/1;5	0.18541	1	SD;inverse;1/5/1;1000	0.27778
2	SD;inverse;1/1/1;1000	0.18524	2	SD;inverse;25/5/1;1000	0.27761
3	SD;inverse;25/5/1;5	0.18496	3	SD;inverse;1/5/1;5	0.27733
3	SD;inverse;1/5/1;5	0.18496	4	SD;inverse;25/5/1;5	0.27716
4	SD;inverse;25/5/1;2	0.18485	5	SD;inverse;1/5/25;5	0.27581
4	SD;inverse;1/5/1;2	0.18485	6	SD;inverse;1/5/25;1000	0.27576
5	SD;inverse;1/1/1;2	0.18468	7	SD;inverse;1/5/25;2	0.2757
6	CD;inverse;1/1/1;5	0.18429	7	SD;inverse;1/5/1;2	0.2757
7	CD;inverse;1/1/1;1000	0.18401	8	SD;inverse;25/5/1;2	0.27565
8	SD;inverse;25/5/1;1000	0.18384	9	SD;inverse;1/1/1;1000	0.27486
8	SD;inverse;1/5/1;1000	0.18384	10	SD;inverse;1/1/1;5	0.2748
9	CD;inverse;1/5/25;5	0.18378	11	SD;inverse;1/1/1;2	0.27447
10	CD;inverse;25/5/1;1000	0.18367	12	CD;inverse;1/1/1;5	0.27189
10	CD;inverse;1/5/1;1000	0.18367	13	CD;inverse;1/1/1;2	0.27172
11	CD;inverse;25/5/1;5	0.18356	14	CD;inverse;1/5/25;2	0.27116
11	CD;inverse;1/5/1;5	0.18356	15	CD;inverse;25/5/1;1000	0.27093
12	SD;inverse;1/5/25;1000	0.18345	15	CD;inverse;1/5/1;1000	0.27093
12	CD;inverse;1/1/1;2	0.18345	16	CD;inverse;25/5/1;5	0.27071
13	CD;inverse;1/5/25;1000	0.18339	16	CD;inverse;1/5/1;5	0.27071
14	SD;inverse;1/5/25;5	0.18328	17	CD;inverse;1/1/1;1000	0.27059
15	CD;inverse;1/5/25;2	0.18305	18	CD;inverse;1/5/1;2	0.27003
16	CD;inverse;25/5/1;2	0.183	19	CD;inverse;25/5/1;2	0.26992
16	CD;inverse;1/5/1;2	0.183	20	CD;inverse;1/5/25;5	0.26975
17	SD;inverse;1/5/25;2	0.18266	20	CD;inverse;1/5/25;1000	0.26975
18	SD;1minus;1/1/1;1000	0.17323	21	SD;1minus;1/1/1;1000	0.26846
19	SD;1minus;25/5/1;2	0.17318	22	SD;none;1/1/1;5	0.26773
20	SD;none;1/5/1;5	0.17273	23	SD;none;1/5/25;1000	0.26762
21	SD;none;1/1/1;5	0.17262	23	SD;1minus;1/1/1;5	0.26762
22	SD;none;1/5/1;2	0.172	24	SD;none;1/1/1;1000	0.267

	Settings	Coverage max 0.25387		Settings	Coverage max 0.34815
23	SD;none;25/5/1;5	0.17189	25	SD;1minus;1/5/25;1000	0.26672
24	SD;1minus;1/5/1;2	0.17172	26	SD;1minus;1/5/25;5	0.26644
25	SD;none;25/5/1;2	0.17155	27	SD;none;1/5/25;5	0.26566
26	SD;1minus;25/5/1;5	0.17144	28	SD;none;1/1/1;2	0.26554
27	SD;none;1/1/1;1000	0.17138	29	SD;none;25/5/1;5	0.26493
28	SD;none;1/5/25;2	0.17121	30	SD;1minus;25/5/1;5	0.26431
29	SD;1minus;1/5/25;1000	0.1711	31	SD;1minus;1/1/1;2	0.26392
30	SD;none;1/1/1;2	0.17104	32	SD;1minus;1/5/1;5	0.26364
30	SD;1minus;1/1/1;2	0.17104	33	SD;none;1/5/1;5	0.2633
30	SD;1minus;1/1/1;5	0.17104	34	SD;1minus;1/5/25;2	0.26246
31	SD;none;25/5/1;1000	0.17071	35	SD;1minus;1/5/1;1000	0.26235
32	SD;1minus;25/5/1;1000	0.17048	36	SD;none;1/5/25;2	0.26212
32	SD;1minus;1/5/1;5	0.17048	37	SD;1minus;25/5/1;2	0.26178
33	SD;1minus;1/5/25;2	0.16992	38	CD;none;25/5/1;5	0.26156
34	SD;1minus;1/5/25;5	0.16942	39	CD;1minus;1/5/1;5	0.2615
35	SD;none;1/5/1;1000	0.16925	40	CD;1minus;25/5/1;5	0.26128
36	SD;none;1/5/25;5	0.16891	41	SD;1minus;1/5/1;2	0.26105
37	SD;none;1/5/25;1000	0.16846	42	CD;none;1/5/1;5	0.26089
38	SD;1minus;1/5/1;1000	0.16779	43	SD;none;1/5/1;1000	0.26049
39	CD;1minus;25/5/1;5	0.1674	43	CD;none;25/5/1;1000	0.26049
40	CD;none;25/5/1;5	0.16689	44	SD;none;25/5/1;2	0.26033
41	CD;none;1/5/1;5	0.16655	44	SD;none;25/5/1;1000	0.26033
42	CD;none;1/5/1;2	0.16633	45	CD;1minus;1/5/1;1000	0.26027
43	CD;none;25/5/1;1000	0.16611	46	CD;1minus;1/1/1;5	0.25982
44	CD;1minus;1/5/1;5	0.16582	47	SD;none;1/5/1;2	0.25971
45	CD;1minus;25/5/1;1000	0.16554	47	SD;1minus;25/5/1;1000	0.25971
46	CD;none;1/5/1;1000	0.16515	48	CD;none;1/1/1;5	0.25965
46	CD;1minus;1/5/1;1000	0.16515	49	CD;none;1/5/1;1000	0.25948
47	CD;1minus;1/5/1;2	0.16453	50	CD;1minus;25/5/1;1000	0.25932
48	CD;none;25/5/1;2	0.1642	51	CD;1minus;1/1/1;1000	0.25903
49	CD;1minus;25/5/1;2	0.16414	52	CD;1minus;1/5/1;2	0.25875
50	CD;1minus;1/1/1;5	0.16403	53	CD;none;25/5/1;2	0.25831
51	CD;1minus;1/1/1;2	0.16324	54	CD;1minus;25/5/1;2	0.25786
52	CD;none;1/1/1;5	0.16218	55	CD;none;1/5/1;2	0.2578
53	CD;none;1/1/1;1000	0.16145	55	CD;none;1/1/1;1000	0.2578
54	CD;none;1/5/25;5	0.16134	56	CD;1minus;1/1/1;2	0.2573
55	CD;none;1/1/1;2	0.16117	57	CD;none;1/5/25;5	0.25629

	Settings	Coverage max 0.25387		Settings	Coverage max 0.34815
56	CD;none;1/5/25;2	0.16038	58	CD;none;1/1/1;2	0.25589
57	CD;1minus;1/5/25;2	0.1601	59	CD;1minus;1/5/25;2	0.25516
58	CD;1minus;1/5/25;5	0.16004	60	CD;1minus;1/5/25;1000	0.25505
59	CD;1minus;1/5/25;1000	0.15999	61	CD;none;1/5/25;2	0.25494
60	CD;1minus;1/1/1;1000	0.15937	62	CD;1minus;1/5/25;5	0.25466
61	CD;none;1/5/25;1000	0.15735	63	CD;none;1/5/25;1000	0.25292
62	SD;normal;1/5/1;5	0.14607	64	SD;normal;25/5/1;5	0.2422
63	SD;normal;1/1/1;5	0.14568	65	SD;normal;1/5/1;5	0.23956
64	SD;normal;1/1/1;1000	0.14461	66	SD;normal;1/1/1;5	0.23838
65	SD;normal;1/5/1;1000	0.14439	67	SD;normal;25/5/1;1000	0.23805
66	SD;normal;25/5/1;1000	0.14433	68	CD;normal;1/1/1;1000	0.23771
67	CD;normal;1/1/1;5	0.14416	69	CD;normal;25/5/1;2	0.23743
68	CD;normal;25/5/1;1000	0.144	69	CD;normal;25/5/1;1000	0.23743
69	CD;normal;1/5/1;2	0.14377	70	SD;normal;1/1/1;1000	0.23715
70	SD;normal;25/5/1;5	0.14343	71	SD;normal;1/5/1;1000	0.23692
71	CD;normal;1/5/1;1000	0.14327	72	CD;normal;1/5/1;5	0.23659
72	CD;normal;25/5/1;2	0.1431	73	SD;normal;1/5/1;2	0.23597
73	SD;normal;1/5/25;1000	0.14304	74	CD;normal;1/5/1;1000	0.23558
74	SD;normal;25/5/1;2	0.14282	75	SD;normal;25/5/1;2	0.23552
75	CD;normal;1/5/1;5	0.14237	76	CD;normal;25/5/1;5	0.23535
76	SD;normal;1/5/1;2	0.14231	77	CD;normal;1/1/1;5	0.2353
77	SD;normal;1/5/25;5	0.14226	78	CD;normal;1/5/25;5	0.23468
78	CD;normal;25/5/1;5	0.14198	79	SD;normal;1/5/25;1000	0.23378
79	CD;normal;1/1/1;1000	0.14074	80	SD;normal;1/5/25;5	0.23339
80	CD;normal;1/5/25;5	0.13945	81	CD;normal;1/5/1;2	0.23182
81	SD;normal;1/1/1;2	0.13906	82	CD;normal;1/5/25;1000	0.23165
82	CD;normal;1/1/1;2	0.139	83	SD;normal;1/1/1;2	0.2307
83	CD;normal;1/5/25;1000	0.13855	84	CD;normal;1/1/1;2	0.22901
84	CD;normal;1/5/25;2	0.13614	85	CD;normal;1/5/25;2	0.22767
85	SD;normal;1/5/25;2	0.13311	86	SD;normal;1/5/25;2	0.22323

Table C.3: Best settings for simple prosody coverage

Best settings for clustered prosody coverage					
Gutenberg			Wikipedia		
	Settings	Coverage max 0.30647		Settings	Coverage max 0.40043

	Settings	Coverage max 0.30647		Settings	Coverage max 0.40043
1	CD;inverse;1/1/1;1000	0.26692	1	CD;inverse;1/1/1;5	0.35425
2	CD;inverse;1/1/1;5	0.2668	2	CD;inverse;25/5/1;1000	0.35368
3	CD;inverse;1/5/25;5	0.26669	2	CD;inverse;1/5/1;1000	0.35368
4	CD;inverse;1/5/25;1000	0.26623	3	CD;inverse;1/5/25;2	0.35334
5	CD;inverse;25/5/1;1000	0.26497	4	CD;inverse;1/5/25;5	0.35311
5	CD;inverse;1/5/1;1000	0.26497	4	CD;inverse;1/1/1;1000	0.35311
6	CD;inverse;25/5/1;5	0.26417	5	CD;inverse;1/5/25;1000	0.35265
6	CD;inverse;1/5/1;5	0.26417	6	CD;inverse;25/5/1;5	0.35219
7	CD;inverse;1/1/1;2	0.26383	6	CD;inverse;1/5/1;5	0.35219
8	CD;inverse;1/5/25;2	0.2636	6	CD;inverse;1/1/1;2	0.35219
9	CD;inverse;25/5/1;2	0.26052	7	CD;inverse;1/5/1;2	0.34865
9	CD;inverse;1/5/1;2	0.26052	8	CD;inverse;25/5/1;2	0.34854
10	CD;1minus;1/1/1;5	0.24508	9	CD;none;25/5/1;1000	0.34351
11	CD;1minus;25/5/1;5	0.24451	10	CD;1minus;1/1/1;1000	0.34236
12	CD;none;25/5/1;1000	0.24417	11	CD;1minus;1/5/1;5	0.34225
13	CD;none;25/5/1;5	0.24406	12	CD;1minus;1/1/1;5	0.34214
14	CD;none;1/5/1;5	0.24383	13	CD;none;1/1/1;5	0.34202
14	CD;1minus;25/5/1;1000	0.24383	14	CD;none;25/5/1;5	0.34191
15	CD;1minus;1/5/25;1000	0.24303	15	CD;none;1/5/1;1000	0.34179
16	CD;1minus;1/5/1;1000	0.24291	15	CD;1minus;1/5/1;1000	0.34179
17	CD;none;1/5/1;1000	0.2428	16	CD;1minus;25/5/1;5	0.34156
18	CD;1minus;1/5/1;5	0.24246	16	CD;1minus;25/5/1;1000	0.34156
19	CD;none;1/1/1;5	0.24234	17	CD;none;1/1/1;1000	0.34122
19	CD;none;1/1/1;1000	0.24234	18	CD;none;1/5/1;5	0.34076
20	CD;none;1/5/25;5	0.24166	19	CD;1minus;1/5/25;1000	0.33928
21	CD;1minus;1/5/25;5	0.24097	20	SD;inverse;1/5/1;1000	0.33836
22	CD;1minus;1/1/1;1000	0.2404	21	CD;none;1/5/25;1000	0.33825
23	SD;inverse;25/5/1;5	0.24017	22	SD;inverse;25/5/1;1000	0.33813
23	SD;inverse;1/5/1;5	0.24017	23	SD;inverse;1/5/1;5	0.33779
24	SD;inverse;1/1/1;1000	0.24005	24	SD;inverse;25/5/1;5	0.33756
25	SD;inverse;1/1/1;2	0.23994	25	SD;inverse;25/5/1;2	0.33688
25	SD;inverse;1/1/1;5	0.23994	25	SD;inverse;1/5/1;2	0.33688
26	CD;none;1/5/1;2	0.23971	25	CD;none;1/5/25;5	0.33688
27	CD;1minus;25/5/1;2	0.2396	26	CD;1minus;1/5/25;5	0.33653
28	SD;inverse;25/5/1;2	0.23937	27	SD;inverse;1/5/25;1000	0.33516
28	SD;inverse;1/5/1;2	0.23937	28	SD;inverse;1/5/25;5	0.33482
29	CD;1minus;1/1/1;2	0.23868	29	SD;inverse;1/5/25;2	0.33471

	Settings	Coverage max 0.30647		Settings	Coverage max 0.40043
30	SD;inverse;25/5/1;1000	0.23857	30	SD;inverse;1/1/1;5	0.33413
30	SD;inverse;1/5/1;1000	0.23857	31	SD;inverse;1/1/1;1000	0.33402
31	CD;none;1/5/25;1000	0.23811	32	CD;1minus;1/5/1;2	0.33379
32	SD;inverse;1/5/25;1000	0.23777	33	CD;none;25/5/1;2	0.33356
33	CD;1minus;1/5/1;2	0.23765	34	CD;1minus;25/5/1;2	0.33333
34	CD;none;25/5/1;2	0.23754	35	SD;inverse;1/1/1;2	0.33276
35	SD;inverse;1/5/25;5	0.2372	36	CD;1minus;1/1/1;2	0.33242
36	CD;none;1/1/1;2	0.23685	37	CD;none;1/5/1;2	0.33219
37	SD;inverse;1/5/25;2	0.23663	38	CD;none;1/1/1;2	0.33173
38	CD;1minus;1/5/25;2	0.23594	39	CD;none;1/5/25;2	0.33048
39	CD;none;1/5/25;2	0.23514	40	CD;1minus;1/5/25;2	0.32956
40	SD;1minus;25/5/1;2	0.22611	41	SD;none;1/1/1;5	0.32236
41	SD;1minus;1/1/1;1000	0.22577	41	SD;1minus;1/1/1;5	0.32236
42	SD;none;1/5/1;5	0.22554	42	SD;1minus;25/5/1;5	0.32202
43	SD;none;1/1/1;5	0.22485	42	SD;1minus;1/1/1;1000	0.32202
43	SD;none;1/1/1;1000	0.22485	43	SD;none;25/5/1;5	0.3219
44	SD;1minus;25/5/1;5	0.22474	44	SD;none;1/1/1;1000	0.32156
45	SD;1minus;1/5/1;2	0.22451	45	SD;none;1/5/25;1000	0.32064
46	SD;none;25/5/1;5	0.22428	46	SD;none;1/1/1;2	0.32042
46	SD;1minus;1/5/25;1000	0.22428	46	SD;1minus;1/5/1;5	0.32042
47	SD;none;25/5/1;2	0.22394	47	SD;none;1/5/1;5	0.32007
48	SD;1minus;1/1/1;5	0.22382	48	SD;1minus;1/5/1;1000	0.31984
49	SD;1minus;1/5/1;5	0.22371	49	SD;1minus;1/5/25;1000	0.31962
50	SD;1minus;1/1/1;2	0.22359	50	SD;none;1/5/25;5	0.31927
51	SD;none;25/5/1;1000	0.22348	51	SD;1minus;1/1/1;2	0.31836
51	SD;none;1/5/1;2	0.22348	52	SD;1minus;1/5/25;5	0.31824
52	SD;none;1/5/25;2	0.22337	53	SD;1minus;25/5/1;2	0.31802
53	SD;none;1/1/1;2	0.22325	54	SD;none;25/5/1;1000	0.31779
54	SD;1minus;1/5/25;2	0.22279	55	SD;1minus;1/5/1;2	0.31767
55	SD;1minus;25/5/1;1000	0.22176	56	SD;none;1/5/1;1000	0.31744
56	SD;1minus;1/5/25;5	0.22165	57	SD;none;25/5/1;2	0.31722
57	SD;none;1/5/25;5	0.22062	58	SD;1minus;1/5/25;2	0.31676
58	SD;none;1/5/1;1000	0.22051	59	SD;none;1/5/1;2	0.31653
59	SD;none;1/5/25;1000	0.22039	60	SD;1minus;25/5/1;1000	0.31561
60	SD;1minus;1/5/1;1000	0.21959	61	SD;none;1/5/25;2	0.31516
61	CD;normal;25/5/1;1000	0.21776	62	CD;normal;1/1/1;1000	0.31401
62	CD;normal;1/5/1;1000	0.21674	63	CD;normal;25/5/1;1000	0.3131

	Settings	Coverage max 0.30647		Settings	Coverage max 0.40043
63	CD;normal;1/1/1;1000	0.21239	64	CD;normal;1/5/1;1000	0.31024
64	CD;normal;1/5/25;1000	0.21045	65	CD;normal;1/5/1;5	0.30807
65	CD;normal;25/5/1;5	0.20953	66	CD;normal;1/1/1;5	0.30533
66	CD;normal;1/1/1;5	0.20919	67	CD;normal;1/5/25;1000	0.30476
67	CD;normal;1/5/1;5	0.20908	68	CD;normal;25/5/1;5	0.3043
68	CD;normal;1/5/1;2	0.20439	69	CD;normal;25/5/1;2	0.3011
69	CD;normal;25/5/1;2	0.20336	70	CD;normal;1/5/25;5	0.3003
70	CD;normal;1/5/25;5	0.20302	71	CD;normal;1/5/1;2	0.2955
71	SD;normal;1/5/1;5	0.19627	72	SD;normal;25/5/1;5	0.29504
72	SD;normal;25/5/1;1000	0.19593	73	SD;normal;1/5/1;5	0.29367
73	CD;normal;1/1/1;2	0.19422	74	SD;normal;1/5/1;1000	0.29104
74	SD;normal;1/5/1;1000	0.19399	75	SD;normal;25/5/1;1000	0.29081
75	SD;normal;1/1/1;5	0.19364	76	SD;normal;1/1/1;5	0.28989
76	SD;normal;25/5/1;5	0.19353	77	CD;normal;1/1/1;2	0.28967
77	SD;normal;1/1/1;1000	0.19227	78	SD;normal;1/1/1;1000	0.28852
78	SD;normal;25/5/1;2	0.19216	79	SD;normal;1/5/1;2	0.28727
79	SD;normal;1/5/25;1000	0.19079	80	SD;normal;25/5/1;2	0.28704
80	CD;normal;1/5/25;2	0.19044	81	CD;normal;1/5/25;2	0.28658
81	SD;normal;1/5/1;2	0.19021	82	SD;normal;1/5/25;1000	0.28612
82	SD;normal;1/5/25;5	0.18816	83	SD;normal;1/5/25;5	0.28406
83	SD;normal;1/1/1;2	0.18599	84	SD;normal;1/1/1;2	0.28064
84	SD;normal;1/5/25;2	0.1789	85	SD;normal;1/5/25;2	0.27229

Table C.4: Best settings for clustered prosody coverage

D Synthesis Scripts

ID	Sentence
dewiki11769_2	deutsch: auch Wojwodina;
dewiki143733_1	Tiengen bezeichnet
dewiki172112_5	Refrain Refrain
dewiki122796_4	Armand hatte noch zwei Schwestern.
dewiki101097_16	Die Hauptrolle übernahm Jeffrey Wright.
dewiki160596_41	sein Nachfolger wurde George Russell.
dewiki603747_23	Das ist genau Russells Paradoxon.
dewiki77166_3	Im Hintergrund die Stadt Coswig.
dewiki164808_1	Kommissär heißt so viel wie-
dewiki64713_2	Tschüß hat seinen Ursprung im Französischen.
dewiki185579_41	dabei werden die Beine abwechselnd gegrätscht).
dewiki632556_5	Jean selbst war der Großvater des Königs Franz.
dewiki75573_19	(Mein Name ist Timothy.
dewiki146201_19	Der Senator billigt die Deutsch Pflicht auf Schulhöfen.
dewiki203675_4	Nach der Schule erlernte er den Beruf des Drehers.
dewiki628876_4	Laurence J.
dewiki183828_4	Bluetooth nutzen.
dewiki170031_14	Am Sockel befindet sich auch ein Porträt Johann Duves.
dewiki107113_5	Man spricht deswegen von einem Zero Knowledge Protokoll.
dewiki103779_13	Unter anderem gründete er eine deutsch chinesische Schule.
dewiki71434_5	Faye Dunaway war zweimal verheiratet.
dewiki22137_9	Den Preis nahm stellvertretend Fr. Jean Marie entgegen.
dewiki155696_56	Am Ende der Saison jedoch verletzte er sich erneut schwer.
dewiki106121_9	Die Einflüsse sind nun indianisch beziehungsweise vom Reggae her kommend.
dewiki179523_28	Carmen wartet verliebt auf José.
dewiki63056_1	Der Begriff Variant bezeichnet
dewiki127475_1	Siehe auch Mannequin.
dewiki95788_5	Um aus Java eine native Methode aufzurufen, muss diese zunächst als" native" deklariert werden.
dewiki184699_380	bei Jean Jacques Rousseau);
dewiki958_296	Danton war am 10. Juli aus dem Ausschuss abberufen worden.

ID	Sentence
dewiki99185_20	Aus dieser Ehe ging Paul Gauguin hervor.
dewiki89558_9	Arpeggio
dewiki616626_20	/ x: Sie zeigen den rechten Arm und das rechte Bein /das war alles/ äh/...
dewiki112040_43	Becker, Jörg;
dewiki70696_78	über Bluetooth kann man eine Verbindung zu anderen Geräten herstellen.
dewiki627166_2	†11. September neunzehn Hundert neunzig bei Alzey) war ein deutscher Jurist.
dewiki64018_19	zwei Tausend drei wurde er von Königin Elizabeth. zum Ritter geschlagen.
dewiki63890_13	Die Aromen der ersten Speise wirken schon vom Teller aus auf den Gourmet.
dewiki70043_4	In der Republik Sacha sind ihre Rechte durch besondere Gesetze geschützt.
dewiki182281_2	So bekommt man schnell einen überblick, in welchem Terrain man sich befindet.
dewiki17538_3	Jedes Jahr im November findet ein Gipfeltreffen der ASEAN Staaten statt.
dewiki120883_19	Auf Deutsch sind;
dewiki88524_12	In diesem Sinne hielten Elizabeth Taylor und George Michael kurze Ansprachen.
dewiki626890_24	Satin Lux Farbe entspricht dem Lux Kaninchen.
dewiki95676_69	Die letzte Szene der Handlung zeigt Andrzej noch schwach, aber genesend im Zimmer des Arztes.
dewiki73484_12	Im gleichen Jahr wurde das erste Solo Album produziert: Fair And Square.
dewiki61114_16	Dadurch kann sie straff gespannt werden und hält selbst böigem Wind stand.
dewiki164008_13	George sprach fließend deutsch.
dewiki775_11	eins von John Cage.
dewiki117974_6	Weitere Bands dieses Sub Genres waren;
dewiki92170_9	Pfuhl, Stahl, stehlen, stöhnen;
dewiki157827_1	Göggingen ist der Name folgender Orte:
dewiki155613_19	Um dies zu tun reicht die Zeit locker aus, die der Spieler benötigt um seinen Einsatz aus dem Portemonnaie zu holen.
dewiki194417_1	Mittweida ist
dewiki9886_28	Demgegenüber gab sich der Punk illusionslos und setzte auf offene Ablehnung der Gesellschaft.
dewiki68589_16	George hatte unterdessen sein Studium abgeschlossen und arbeitete auf einer Ranch.
dewiki66269_3	Sie war von der Außenwelt schwer zugänglich und gilt als die ärmste Provinz Chinas.
dewiki627583_1	Gouvernement
dewiki70748_4	Um die Mitte des Jahrhunderts kam es deswegen zu Auseinandersetzungen mit den Einheimischen(natives).
dewiki62182_1	Tachymeter steht für:
dewiki8185_26	Bluetooth gilt nur dann nicht mehr als sicher, wenn der Code zu kurz gewählt ist(etwa vier;

ID	Sentence
dewiki187174_7	In der Spitze hatte sie sechs Tausend Abonnementen.
dewiki155079_22	Die Instrumente wirken ungleich, aber stets genial, je später, desto individueller.
dewiki141108_7	Im Club treten bekannte und auf.
dewiki621562_18	Jane Fonda wurde im Jahr 1990 für die Goldene Himbeere nominiert.
dewiki82780_1	Radio China International(;
dewiki122909_10	Pjotr Tschaikowski war einer der ersten Komponisten, der dieses Instrument auch im Orchester einsetzte.
dewiki68000_24	Beispiel: Hui!
dewiki10885_41	b den Phonemen/ b /(Erbe) und / p /(Erbse), oder v den Phonemen / f/(Vater) und/ v/(Vulkan).
dewiki127724_6	Im 16. Jahrhundert wurde dieser Name noch Xavier geschrieben, aber, später ausgesprochen.
dewiki617807_21	Dort verletzt er Cynthia mit einem Messer schwer.
dewiki95676_63	Sie sucht den Mediziner auf und möchte von ihm wissen, ob Andrzej leben oder sterben wird.
dewiki116201_4	Noch heute sieht man Queen Elizabeth. bei feierlichen Anlässen mit der königlichen Kutsche vorfahren.
dewiki12319_14	Jahrhundert monophthongiert(während er im Bairischen und im Alemannischen bis heute auftritt).
dewiki3135_53	(Königin Elizabeths Schlüssel.).
dewiki64177_111	Darunter war erstmals auch ein Rundfunk Journalist.
dewiki160892_30	es ist die Enterprise...
dewiki990_45	Der Freitag hat seinen Namen ebenfalls von der Göttin.
dewiki93967_8	Siehe auch: Snob
dewiki8562_196	Framework.
dewiki143816_20	Jean-Claude van Damme ist zum fünften Mal verheiratet.
dewiki92950_30	Nur einen Monat später heiratete er seine jetzige Frau Estelle Cruyff, eine Nichte von Johan Cruyff.
dewiki130391_22	Durch ihn lernte er Paul Gauguin kennen.
dewiki629466_3	An der Seite von Jean Richard spielt sie dessen Frau Christine, die sich zeitweise in einen Panther verwandelt.
dewiki192010_35	Diese Ablehnung überwand Audrey Richards durch ihre stetige Arbeit.
dewiki10282_7	(deutsch:" Helft Mir!").
dewiki77953_1	Freia ist
dewiki187258_10	Alain Robert: Mit nackten Händen.
dewiki608072_16	Schließlich wurde es von Simon and Schuster publiziert.
dewiki634513_7	Zwerg Seepferdchen werden nur zwei Zentimeter lang.
dewiki170944_3	Der Club spielte von Beginn an in der ersten rumänischen Liga

ID	Sentence
	und änderte sein Namen ein Jahr später in(Central Armee Sport Club).
dewiki115341_1	Prinz Johann von Frankreich(frz. Jean de Berry;
dewiki203248_5	An ein bestimmtes Material ist Satin nicht gebunden.
dewiki102353_1	Bayreuther Festspiele
dewiki86527_68	Die ausschlüpfenden Jungen erfahren keine Metamorphose;
dewiki19829_121	Ein Jüngling musste der Familie der zukünftigen Frau Pferde geben.
dewiki195007_20	(Anekdote Heinz Damian, Club Kassier).
dewiki607080_5	mit Paul Rutherford und John Stevens).
dewiki18981_4	Der native Name der Sprache ist(oder) für die geschriebene Sprache und sechzig(oder) für die gesprochene Sprache.
dewiki171706_20	Ein Jahr später wurde das Vorwerk in eine Erbpacht umgewandelt und Johann George Jahn wurde der erste Pächter.
dewiki445_23	Die Prähistorie, also die Vorgeschichte, umfasst den Zeitraum vom Beginn der Menschwerdung bis zur Einführung der Schrift.
dewiki148930_2	Wow) ist ein Treffen nordamerikanischer Indianer.
dewiki195224_8	Der Sitz des Ministerium war in Strausberg bei Berlin.
dewiki84501_1	Der Begriff Jus bezeichnet:
dewiki154662_21	And Back veröffentlicht.
dewiki204005_10	Das Libretto schrieb Samuel Humphreys.
dewiki110783_4	Lampions werden gern bei Festlichkeiten außer Haus verwendet.
dewiki73816_109	Im Untergeschoss hat der traditionsreiche Club zu Bremen seine Räume, er ist hervorgegangen aus der Gesellschaft Museum.
dewiki63131_41	bis zum Jahr 2000 hatte Madrid acht Champions League Titel (sechzehn Punkte) und zwei UEFA Pokal Titel(zwei Punkte) gewonnen.
dewiki113592_21	Nach der Gründung der Tschechoslowakei kümmerte sich der aus privaten Mitteln finanzierte Klub Tschechischer Touristen um das Bauwerk.
dewiki138708_44	Für das Bellen muss der Kehlkopf jedoch relativ groß sein.
dewiki82957_42	Alternierende Verse werden auch als jambisch oder trochäisch bezeichnet.
dewiki182796_5	Turner war außerdem Wegbereiter für die klassische Soul Musik, die er in der Ike and Tina Turner Revue einem großen Publikum darbot.
dewiki8392_7	bei Dolmetsch).
dewiki175161_41	Einige weitere Szenen wurden geschnitten, um die Handlung zu straffen (auf eine Dauer von circa 113 Min./ PAL)
dewiki19853_17	Im Jahr darauf wurde sein Sohn Jean geboren.
dewiki8726_1	Der Begriff Hai bezeichnet: Siehe auch: Hey
dewiki60040_65	Er hat es sich zur Aufgabe gemacht Roy an die Spitze zu bringen.
dewiki95676_68	Er versichert, dass Andrzej keine Chance auf ein überleben hat und sterben wird, dabei beruft er sich auf Gott als seinen Zeugen.

ID	Sentence
dewiki623791_7	Er war verheiratet mit Rouge Ackermann.
dewiki89591_9	Das ebenfalls weit verbreite S / Protokoll verwendet dagegen Zertifikate und ist deshalb grundsätzlich nicht kompatibel zu.
dewiki9503_48	Siehe Tachometer.
dewiki178778_11	Die Lieder haben oft komplexere Struktur als das typische Strophe Refrain Schema.
dewiki75238_1	Der Begriff Teamwork bezeichnet
dewiki61392_50	Populär in den Städten ist der Reggae.
dewiki183478_15	Ferner glaubte man seitens der Regierung Informationen über eine angebliche Verschiebung deutscher Truppen in Richtung Tschechoslowakei zu haben.
dewiki107250_42	Die Ernte kam sehr früh heim.
dewiki108230_16	Zudem steht bereits die Roadmap für die Version vier fest.
dewiki96189_30	League zu gehen.
dewiki168852_6	Der amerikanische Architekt Richard Meier hat einen lichten, offenen Bau mit zwei großen Sälen, zwei Kabinetten und einem Souterrain geschaffen.
dewiki9950_1	Gateway ist
dewiki81290_1	Moroder ist der Name von
dewiki83121_22	In größter Bedrängnis Chinas Volk.
dewiki193114_1	Jeremias.
dewiki5262_4	auf Pfählen).
dewiki22768_23	Champions League Sieger 2001;
dewiki72304_23	Die kreative Leistung Jil Sanders ist unbestritten.
dewiki153573_22	Und ich dann so:" Dankeschön.
dewiki602642_10	Für und sollen in Zukunft native Treiber zur Verfügung gestellt werden.
dewiki5699_5	Ein jüdischer Witz definiert und illustriert zugleich Chuzpe so:
dewiki174015_117	Offen gelassen wird die Frage, ob nun auch Ron und Hermine ein Paar sind.
dewiki17363_70	Wenig später übernahm der Deutsche Fußball Bund() die Regel aus Bayern, kurz darauf folgten die Europäische Fußball Union(UEFA) und der Weltverband FIFA.
dewiki8185_35	Jedoch muss der Angreifer die Bluetooth Adresse eines verbunden Bluetooth Moduls kennen.
dewiki70410_8	Zuletzt arbeitete er als Pförtner.
dewiki127180_2	Siehe Enveloppe(Mathematik).
dewiki9435_15	Alain Prost dürfte in die F1 Geschichte als der Rennfahrer eingehen, der wie kaum ein zweiter den Typus des Analytikers hinter dem Lenkrad verkörperte.
dewiki188434_50	(Refrain:) Seid stolz!
dewiki92344_33	Bluetooth wird kommen, daran besteht bei vielen Herstellern

ID	Sentence
	kaum noch ein Zweifel.
dewiki117447_70	Sie ist außerhalb Chinas kaum bekannt.
dewiki612260_3	Der Titel des Stückes wird im französischen Original mit einem" m" geschrieben(Dom Juan);
dewiki13667_11	Der Cancan galt als wild, anstößig und obszön.
dewiki164659_10	Der nächste Bahnhof befindet sich in Pasewalk.
dewiki2610_11	Ab Lyon fließt er von Nord nach Süd.
dewiki19415_1	Ilja Michailowitsch Frank(russisch;
dewiki70784_52	Peter Pichler, Akkordeon;
dewiki107113_6	Insbesondere ist das Protokoll perfekt zero knowledge.
dewiki69741_19	Sie können sehr gut klettern, hoch springen und passen durch alle Löcher, durch die auch ihr Kopf passt;
dewiki16865_54	Berühmt sind die vielen alten Gobelin Wandteppiche, die in einigen Räumen hängen.
dewiki75249_58	grus(Westlicher Kranich) and Grus g.
dewiki88850_19	George Green, Hermann Grassmann).
dewiki97064_7	zwei Tausend zwei zog er sich in den Ruhestand zurück.
dewiki203514_35	Elizabeth von Arnim kehrte mit ihren fünf Kindern nach Großbritannien zurück.
dewiki2893_89	Auf Chinesisch heißt Struwwelpeter übrigens-;
dewiki63789_3	Auch unter Machête bekannt.
dewiki627735_21	Nach seinem Rücktritt zog sich Wright aus der Politik zurück.
dewiki605295_4	Anschließend erhält jeder Spieler pro Chip eine Karte.
dewiki96930_2	Die Stadt Höchstadt selbst ist kein Mitglied der.
dewiki187751_42	Zudem sollen 3D Objekte wie Bäume und Steine über das Terrain verteilt werden können.
dewiki21197_187	Ebenfalls von Bedeutung ist der Queen Elizabeth Park.
dewiki114825_5	Damit war es der erste wirtschaftlich nutzbare synthetische Kautschuk.
dewiki13921_16	Die Änderung hat das Forum Train Europe() beschlossen, dem auch die Bahnen im deutschsprachigen Raum gehören.
dewiki188235_15	Sie hatten drei Kinder Elizabeth Rose, Richard und Robert.
dewiki84703_3	Er trat auch unter dem Pseudonym Bjarne auf.
dewiki164787_19	Mike Pflüger nahm seinen Platz ein.
dewiki65298_27	Dabei handelte es sich um eine deutsch tschechische Koproduktion.
dewiki106572_5	Siehe auch: Blei(,)- oxid
dewiki61818_55	Viele hielten ihn schlicht für' einen blöden Macho'.
dewiki176283_6	Ein wöchentliches Engagement im Club folgte.
dewiki635286_18	Die Hauptrolle spielte Sean Connery nach zwölf Jahren Pause von der Rolle.

ID	Sentence
dewiki610654_35	Simon soll immer Klaras Freund bleiben.
dewiki16082_7	für einzelne Herren ist ein Club Besuch am teuersten.
dewiki1847_7	Hochdeutsch Niederdeutsch
dewiki522_33	Sein Stil war neu und galt in weit über das wissenschaftliche Milieu hinausgehenden Kreisen als spektakulär.
dewiki636619_18	Er verlässt den Club mit Liz.
dewiki74042_5	Botanik und Paläontologie.
dewiki2240_13	vergleiche hierzu Chinesisches Orakel.
dewiki102613_85	(dt. jetzt oder nie;
dewiki162886_11	Es existieren zwei Fischerei Betriebe und eine blühendes Charter Fischerei Geschäft.
dewiki953_228	statt des a kann ein ä stehen.
dewiki11584_141	Der Putsch misslang.
dewiki199973_20	Es klirrten die Becher, es jauchzten die Knecht;
dewiki630620_11	Im Jahre 1504 stand noch eine Kapelle auf einer zum Dorf gehörenden Wiese and der Elster.
dewiki4939_69	vor George Lucas'.
dewiki626200_6	Eine Version des Symbols wird als Trademark von der Firma
dewiki96308_15	Das Saisonziel ist die Qualifikation für einen internationalen Wettbewerb.
dewiki11311_29	ltd., Singapur;
dewiki187524_169	drei anknüpft.
dewiki74303_10	Francisco Javier
dewiki162208_10	Gerald Allen and Norbert J.
dewiki627765_3	Meister wurde der Deutsche Rugby Club Hannover(Hannover).
dewiki118737_21	Im Gesamtklassement wurde er Fünfter.
dewiki9834_71	Ihnen ist der Freitag gewidmet.
dewiki84097_4	Noahs Arche soll hier gelandet sein.
dewiki134726_11	Bis dahin wurde die Straße meistens erst ab Pfingsten geräumt.
dewiki16953_3	Gehört zu den Olympiern.
dewiki96296_5	Konkret sind dies Berufe im medizinisch pflegerischen, im sozialen und auch im künstlerischen Bereich.
dewiki179523_66	Carmen bleibt mit José zurück.
dewiki148635_52	Luftballons werden in unterschiedlichen Bereichen und zu unterschiedlichen Zwecken eingesetzt.
dewiki131127_6	Für die schulischer Weiterbildung sorgt das am Ort ansässige Lyon College.
dewiki1898_299	de da null de da null de da null;
dewiki363_42	Danach stieg Charles noch einmal selbst alleine auf.
dewiki18793_23	Auf Betreiben des späteren Paul. ernannte Paul.

ID	Sentence
dewiki92562_9	Vergleichbar ist auch das Bauernfrühstück.
dewiki8649_21	Es ist Ebbe auf dem Konto, Die Staatskasse zeigt Ebbe.
dewiki135192_10	Matthew J.
dewiki180887_1	À point(frz.)(auch englisch oder rosa;
dewiki67271_39	Im gleichen Jahr brachte er das Album I Am What I Am heraus, das sich mehr als eine Million mal verkaufte und mit Platin ausgezeichnet wurde.
dewiki623591_77	Auch über das anstehende amerikanische Remake.
dewiki176927_15	Yale, zwei Tausend vier
dewiki105805_19	Einige Schüler gründen diesen Club neu.
dewiki105314_8	Josef Pröll ist der Neffe von Erwin Pröll.
dewiki109825_11	Aber auch aus Vertrag kann jemand ersatzpflichtig werden.
dewiki624619_13	Nach ihrer Heirat wurde Henriette Mitglied der.
dewiki1012_30	Chopin bemühte sich zunächst, weiter auf dem Klavier unterrichtet zu werden.
dewiki14485_39	Einer, der die Damen mit viel Takt anquatscht.'
dewiki132782_18	Diese verursachten ein Verkehrschaos im Stadtzentrum.
dewiki110810_27	Freitag stürzte Fette mit Rückendeckung der Metall und wurde neuer Vorsitzender des.
dewiki191731_2	Die Idee zur Serie hatte Christopher Crowe.
dewiki610255_14	Schon bald wird Justine anonym verlegt.
dewiki127819_25	Sie waren die erfolgreichsten Trainer der" Irons".
dewiki73079_17	Durch ihn wurde ein großer Teil Chinas für die westliche Wissenschaft erschlossen.
dewiki156424_9	Für die Rolle der Judy erhielt sie eine Oscar Nominierung.
dewiki134725_1	Als Vertrag von Oslo werden verschiedene Abkommen bezeichnet, die in der Stadt Oslo geschlossen wurden:
dewiki68649_27	Es sollte die Antwort des auf den Beat Club der sein.
dewiki166833_65	Es waren die Obere Pforte wie die Untere Pforte.
dewiki607624_8	Noch im selben Jahr erfolgte seine Habilitation für das Fach Neurochirurgie.
dewiki12444_26	Dieses Milieu zog Banditen aller Art geradezu magisch an.
dewiki604129_1	Henriette steht für:
dewiki80149_8	Er systematisierte die Resultate von George Boole.
dewiki140148_1	Der Berliner Fußball Club Viktoria ein;
dewiki10944_60	Diese Lieder sind gewöhnlich mit" ö" oder"(ö)" gekennzeichnet.
dewiki124481_2	Sie liegt nördlich von Lörrach.
dewiki126222_6	Zu seinen Lehrern zählten Schostakowitsch und Prokofjew.
dewiki199777_17	Seine Kanzlei betreut unter anderem Guantánamo Häftlinge.
dewiki165616_45	Klicker sind kugelförmige, geschliffene Achate.
dewiki111713_8	bei Bluetooth, und W verwendet.

ID	Sentence
dewiki636336_1	George Taylor ist der Name von
dewiki131892_221	Vier Alleen führen geradeaus auf den Pavillon zu.
dewiki186577_105	Sein Plan, sich Rheine Untertan zu machen, gelingt.
dewiki125393_9	Jacques Becker ist der Vater des französischen Regisseurs Jean Becker.
dewiki621266_3	Verheiratet war er mit Cathrine, geb. Schram.
dewiki5515_113	Dieser führt in der Regel zu Aids definierenden Erkrankungen (Klassifikation C, siehe Aids).
dewiki118630_1	Als Punch bezeichnet man
dewiki65832_1	George Clinton ist der Name folgender Personen:
dewiki81576_10	Bis dato führte er auch das Gesamtklassement an.
dewiki153142_2	Herausgeber ist der Verein für Sprachpflege ein().
dewiki615797_5	Sein Ritter Titel der Loge war" Leonardo da Vinci".
dewiki191346_11	er unterstützte allerdings mit Überzeugung Roosevelts Politik des New Deal.
dewiki100291_11	m ü) noch in einer Entfernung von vierzehn;
dewiki195232_16	Darauf hin verlieh man ihm am 2. März s.
dewiki68911_9	So fotografierte sie zum Beispiel Charles Darwin.
dewiki138226_28	(Das Bild ist übrigens geradezu genial...
dewiki202874_8	Im selben Jahr heiratete er Anna Hausmann.
dewiki189504_2	Ferner ist Train die Bezeichnung für
dewiki6508_41	Neun Monate nach Titus' Geburt starb Saskia.
dewiki8171_2	Im Jargon lautet die Abkürzung dafür.
dewiki132121_5	Beer, Hans de Beer, Hans de
dewiki198181_7	Ohne diese kann man auch von einer Hommage sprechen.
dewiki157283_1	Georges J.
dewiki12067_11	Sie blieb über zwei Tausend Jahre lang ungelöst.
dewiki176107_13	Außerdem fehlt auch die native Unterstützung der nächsten Internet Protokoll Generation sechs.
dewiki136567_4	Polyester und Satin).
dewiki15563_28	Außerdem ist er Mitglied der Sachbuch Jury der Süddeutschen Zeitung.
dewiki142187_13	Zuletzt war Müller im Dachau.
dewiki71652_2	Das Image passt auch auf eine Mini.
dewiki605237_22	Der Maharadscha lässt ihn von seinen Tigern zerfleischen.
dewiki172195_16	Er hat eine jüngere Schwester, Diane.
dewiki122333_1	Artikel zur Geschichte Chinas hier einordnen.
dewiki607602_1	Topos(der, Plural Topoi;
dewiki203225_16	Sein ebenfalls im Jahr zwei Tausend veröffentlichtes Album Wow!
dewiki167749_25	b gleich b and;
dewiki20027_3	Sie wird dann Bridge oder Bridge genannt.

ID	Sentence
dewiki5869_138	Sein zehntes Triptychon, Amazonen, blieb unvollendet.
dewiki174106_24	Wie geplant bombardieren sie Tokio und gehen dann auf Kurs Richtung China.
dewiki98708_15	Regie: Xavier Koller.
dewiki10960_27	Pfauen sind polygame Vögel.
dewiki602416_23	Weiterhin befindet sich dort ein Kanal Bassin.
dewiki120584_90	Ergebnisse stets aus uruguayischer Sicht
dewiki168032_7	Die Renoncen(Füchse) tragen ein Band in grün weißgrün.
dewiki188541_1	Joe Armstrong(;
dewiki318_18	Bill Murray hat sechs Kinder aus zwei Ehen mit Margaret Kelly und seiner derzeitigen Frau Jennifer Butler.
dewiki156341_15	Daraufhin beschlossen die Amerikaner, sich nach Fort George zurückzuziehen, sahen sich aber eingeschlossen.
dewiki628959_10	Noch galt er als Experte für die Landwirtschaft und noch unterstütze er Chruschtschow.
dewiki14789_18	Da China seine Ansprüche auf den Raum nicht aufgegeben hat, betrachtet es diesen formellen Akt als illegal.
dewiki169675_3	Sie liegt an der Nahe.
dewiki338_251	Nachdem er aufhört zu dribbeln und noch in der Bewegung;
dewiki113578_53	Er wurde entlarvt und ausgepeitscht.
dewiki631205_14	Anstelle des Thunfischs kann auch Lachs verwendet werden.
dewiki81574_26	Mit der auf Eduards.
dewiki133569_23	Ich wollte wieder in Richtung Pforte gehen.

Table D.1: Synthesis script