Preface

1 Introduction

1.1 Word frequency

Through the ages scholars have, for many different reasons, sought to acquire an understanding of the vocabulary of individual languages. The study of word frequency has at times received their special attention. The fact that words occur in language with varying frequency is scarcely a recent discovery; some occur fairly regularly in both spoken and written language, while it may take some searching to find examples of the use of others. Lexicographers are well aware of the fact that it is fairly easy to obtain information on some words but may be difficult to obtain reliable information regarding others. This is evident in Figure 1 on the following page, which shows the relationship between the number of word forms and the number of individual words of text in the survey of word frequency described in this book. The survey included texts totalling some 500,000 running words. The graph shows that the 100 most common word forms occurred in almost 300,000 instances. Such a relationship is well known from other surveys.

Actual research on the frequency of words, however, is a relatively recent phenomenon. Interestingly enough, there were two different groups who became interested in word frequency at about the same time, stenographers and educators. The first investigations were made by stenographers, who realized the importance of knowing which words occurred most often in a language. Among the best known examples is the extensive research carried out by the German F. W. Kaeding, who counted the frequency of words in a corpus of approximately ten million words in 1898. Educators first became interested in word frequency in connection with the education of children in language arts. A well known example is the research of the American educator E. Thorndike on the frequency of words in American English, which was first published in 1921. Of Thorndike's later research it is his investigation which was published in 1944 in the book co-authored by Lorge, The Teacher's Word Book of 30,000 Words, which probably became most widely known. Interestingly enough, Thorndike's research on word frequency was later to form the basis for extensive lexicographical work in the US (Landau 1989).¹

Reports of the research by Thorndike and others made their way to this country and inspired the Icelandic research into word frequency carried out by school principal Ársæll Sigurðsson. As early as 1938 he wrote an article in Menntamál: Tímarit um uppeldis- og freðslumál (Educational Affairs: A periodical concerning matters of education and instruction) explaining the need for research on word frequency for teaching of language arts, especially in order to discover the most commonly occurring words in the language or "basic words".

¹ The excellent English-Icelandic dictionary, Ensk-Íslensk orðabók, published by Orn & Órlygur in 1984 is based originally on a dictionary of Thorndike's.
As he put it (Ársæll Sigurðsson 1938:101):

It seems obvious how much assistance it would be to the teacher to keep these "basic words" in mind in his instruction, even though they were limited to 500–1000 words. It could also prove useful to ascertain, if possible, what words the children themselves use most often when they are describing something orally or in writing. This represents the history of the development of the living language of each individual and the form which expression takes at various stages in life.

In the same article Sigurðsson mentions that he has already begun work on such an investigation. The results were subsequently published in Menntamál in 1940. Sigurðsson's investigation was remarkable in many respects and will be examined in more detail later.

Interest in word frequency is, however, by no means limited to scholars in education. Mention was previously made of the interest shown by stenographers in word frequency. Similar interest appears among scholars of information and data processing, who have extensively investigated the quantitative aspects of languages, particularly with the aim of making it easier to devise suitable methods of reducing the space needed for computer storage of written language and to find methods which enable its secure storage. In recent years information on word frequency has also been necessary for the compilation of software such as spelling checkers (McIlroy 1982).

Linguists have perhaps failed to devote as much effort to research on word frequency
as might have been expected but interest in such research has increased in recent years. Doubtless one major factor in this is the increased access by scholars to computerized data after word processing by computer became so widespread. As matters now stand it is relatively easy for almost anyone to acquire a substantial corpus of texts in computerized format.

With the advent of computer technology it has become easier to carry out research on word frequency. This is by no means to say that the computer has relieved all the burdens of scholars, as will become apparent later when the work entailed in this project is described. Professor Baldur Jónsson, who investigated the frequency of words in the novel Hreîðrîð (The Nest) in 1973–1974, was the first to carry out computer-assisted research on word frequency in Iceland. Since that time numerous similar investigations have been carried out. The main ones will be mentioned here, and the organization of the survey upon which the *Frequency Dictionary of Icelandic* is based, along with the manner in which it was conducted, will be described in detail.

Preparations for the survey which is presented in this book first began in 1984 when the Institute of Lexicography began compiling a collection of words for a spelling check program which IBM in Iceland proposed to market. As was mentioned previously, any such word collection must be compiled with regard to word frequency. The fact that approximately 100 word forms account for almost one-half of all the running words in normal written texts affects the way in which such a word collection should be organized, in order to make the search for errors as efficient as possible. One aspect of the preparation of the program was a special survey of word frequency. Some of the texts which were surveyed at that time were also used in preparing this book.

Soon afterwards it was decided that the Institute of Lexicography should acquire a collection of computerized texts to be used in work on the historical dictionary which is in progress at the Institute. The idea was that such a text corpus could be used to search for examples of word usage. The text corpus has grown considerably in the past few years and proved to be as useful as was expected. It was subsequently decided to organize a comprehensive survey of word frequency which would aim at describing in detail the frequency of words in modern Icelandic. The board of directors of the Institute agreed on the project in April of 1985. In so doing the board aimed primarily at supporting research on modern Icelandic language usage.

### 1.2 Earlier research

#### 1.2.1 Studies in other countries

As was previously mentioned, the American educationalist Thorndike was among the first to investigate word frequency. While it is not our intent to discuss research in other countries on the word frequency of various languages comprehensively, two foreign research projects should be mentioned as the present survey was designed with regard for them.

In the 1960s an important investigation into word frequency was begun at Brown University in the US which has since become identified with that Institution. The research is especially noteworthy for four reasons:

1. The texts used in the survey were chosen from different subject areas.
2. Statistical methods were used to choose the texts within each subject area.
3. All the words in the text received an exact grammatical analysis.
4. The researchers at Brown University have authorized other parties to use their text collection ("the Brown Corpus") with the result that scholars have been able to make use of the corpus for various investigations.

The Brown survey extended to a total of fifteen subject areas, including news, religion, poetry, science and humour. Several samples of text were taken from each subject area, each of them some two thousand words in length. The Brown Corpus totals one million running words, and is thus comprised of five hundred individual samples of text. The weighting of the various subject areas varies, however.

All the samples of text were from the year 1961 and were chosen at random after the weighting of the various subject areas had been decided upon.

The actual grammatical analysis took place somewhat later, during the years 1970–1978. The results of the research were published in their final form in 1982 (Francis og Kučera 1982).

The Brown research examined American English. In 1970 a similar project was begun at the University of Lancaster in Britain which aimed at making a comparable survey of British English. The project continued for a number of years and was finally completed in Norway under the direction of Stig Johansson and Knut Hofland. They published their main results in a book in 1989 (Johansson og Hofland 1989a, 1989b). This research also included one million words of running text, comprised of five hundred samples taken from fifteen subject areas, as it was intended to allow for comparisons between American and British English. In the wake of this research a number of noteworthy studies of the mechanical analysis of texts have been published. The Brown Corpus was analysed on a mechanical basis in its time but the results were not very successful. Much better results were obtained in the analysis of the British corpus as described by Garside, Leech and Sampson (1987).

The methodology of these two studies has directly or indirectly influenced many other frequency surveys. Among them are, for instance, the survey carried out by Sture Allén on word frequency in Swedish newspapers (Allén 1970, 1971) and the study by Gunnel Engwall on word frequency in French novels during the years 1962–1968 (Engwall 1984).

All of the studies mentioned above served as a source of inspiration for the survey which is presented here. We thus decided in the beginning to choose texts from various subject areas (which will be described later in more detail) and to analyse the inflections of all the words in detail. The methodology of this survey will be described in more detail later.

### 1.2.2 Icelandic studies

#### 1.2.2.1 Research by Ársæll Sigurðsson

The school principal Ársæll Sigurðsson was the first to study word frequency in modern Icelandic. He published his results in an article entitled Algengustu orðmyndir málsins og stafsetningarkennslan (The most common word forms in the language and spelling instruction) in Menntamál in 1940. In his study Sigurðsson attempts to determine the most common vocabulary of Icelandic, "with the primary aim of finding ways to make spelling instruction more meaningful and realistic than in the past, and at the same time more likely to prove effective." (Ársæll Sigurðsson 1940:9–10). Sigurðsson chose his material, in accordance with his purpose, from the following subject areas: children's essays, adults' letters, school readers, natural science, history and geography.
The main results of Ársæll Sigurðsson's research are given below. The total number of running words turned out to be 100,227, of which 3,530 were proper nouns (3.52%) and 49 dates (0.5%). There were 13,636 different word forms.

The survey revealed in particular that the one hundred most common word forms, which comprise only 0.73% of the total number of word forms, occur in 50,263 instances, which accounts for 52.01% of the running words. It also showed that the majority of the word forms, or 60.48%, occurred only once.

Ársæll Sigurðsson himself had the following to say regarding the results (1940:19–20):

The comparison shows clearly that there are relatively few word forms which comprise the majority of texts, and that, on the other hand, the majority of individual word forms comprise only a small portion of text. [...] This thus proves that the same is true of Icelandic as of other civilized languages, that written language (and probably spoken language as well) is comprised of a few word forms, which are repeated often, and of a large number of word forms which are seldom, if ever, repeated.

1.2.2.2 The study by Baldur Jónsson

The study carried out by Ársæll Sigurðsson described above was, as might be imagined, done "by hand", as no computers were available to assist at the time. Computers had, however, arrived on the scene by the beginning of the 1970s when Baldur Jónsson, Björn Ellertsson and Sven P. Sigurðsson began their survey of the novel The Nest (Hreïðrîð) by Icelandic writer Ólafur Jóhann Sigurðsson. Work on the project was carried out in 1973–74 and the results published in three frequency lists in 1975 entitled Tönni orða í Hreïðrînu 1–3 (The frequency of words in The Nest 1–3), and subsequently five years later in a discussion of the survey and further results with the title Tölvíkðönnun á tönni orða og staфа i Íslensku textu (A computerized survey of the frequency of words and letters in an Icelandic text). Another result of this project was the Orðsthölykill á Hreïðrînu (A concordance of The Nest) which was published in 1978.

The frequency survey may be divided into two sections: on the one hand, a survey of word frequency, and, on the other hand, a survey of the frequency of symbols, the relationship between the length of words and frequency, the average length of word forms, the average frequency of word forms, etc. The results of the frequency survey were published in the three frequency lists previously mentioned (Baldur Jónsson 1975). The first contains the word forms in alphabetical order according to their initial letter, the second word forms in alphabetical order according to their final letter and the third word forms arranged in order of decreasing frequency. The statistical results were subsequently published in the discussion of the survey (Baldur Jónsson et al. 1980:55–126) in the form of lists and tables. Among the information presented is a comparison of the frequency of the one hundred most frequent word forms in The Nest (Hreïðrîð) and in the text corpus used by Ársæll Sigurðsson, the frequency of letters, the arrangement of vowels and consonants in the most common word forms and written words, the frequency of digrams and a comprehensive summary of the frequency of word forms and symbols.

The number of written words turned out to be 53,226 and the number of word forms 11,341, with the average frequency of word forms 4.70.
1.2.2.3 Other Icelandic Investigations

There is little point in dealing extensively with other frequency surveys of Icelandic (some of them are described in detail in an article by Friðrik Magnússon in the periodical Orð og tunga (Word and language), Volume 1), but mention should be made of several which have been discussed in printed sources.

In 1979 Indriði Gíslason and Sigríður Valgeirsdóttir published the results of a survey on the frequency of past subjunctives in the periodical Íslenskt mál (The Icelandic language). As was the case with Ársæll Sigurðsson, Gíslason and Valgeirsdóttir were prompted by educational aims in carrying out their survey (Indriði Gíslason and Sigríður Valgeirsdóttir 1979:107–8):

    Teachers and student teachers were interested in investigating the frequency of past subjunctives, as it is well known that considerable time is spent in language arts in teaching the proper written forms of such verbs. Students are especially prone to err in using the past subjunctive of strong verbs and, if spelling textbooks and departmental examinations are any indication, considerable emphasis appears to be placed on word forms of this type without consideration of their actual frequency.

The material used in the survey was, on the one hand, forty issues of daily newspapers of the year 1975, totalling 49,371.7 column centimetres, and, on the other hand, sixteen issues of four daily newspapers of the year 1925, totalling 7,673.8 column centimetres. The number of running words as calculated from the number of column centimetres was 738,106.9 from 1975 and 114,723.4 from 1925.

The main conclusions were that the verbs vera (to be), hafa (to have) and verða (to become) occurred most frequently in the past subjunctive and together comprise 57.6% of all the verbs in the past subjunctive in 1925 and 63% in 1975. Gíslason and Valgeirsdóttir also concluded from their results that the use of the past subjunctive hardly appears to be declining.

Finally, it should be mentioned that in recent years a number of studies have investigated the vocabulary of the Eddic poems (Baldur Jónsson 1990), the Bible (Baldur Pálsson 1990) and the Icelandic sagas (Eiríkur Rúgnvaldsson 1990).

2 Concepts

Several concepts which were used in the discussion of the previous chapter may have appeared unfamiliar to the reader. Mention was made of running words and word forms, words and inflected forms, without offering any further definitions. Before proceeding further an attempt will be made to define some of the main concepts which are used in studies of word frequency.

The concept of word can have a fair number of meanings and it is necessary to distinguish between the various meanings which it can represent. Readers are doubtless familiar with the practice of counting the number of words in telegrams. There is generally no doubt about how many words are involved, but the manner in which they are written can often determine how they are counted. People are often in doubt as to whether various conjunctions and short words should be written separately or together with others. It is, for instance, common practice to write allt of, einskonar instead of eins konar etc. Furthermore, the proper way to write some words is often a matter of opinion. The word Islendingasögur is most often written as one word but also occurs as two words: Islendinga sögur. The matter becomes even
more difficult when we consider words which are comprised of units other than letters of the alphabet. Is \( \text{H}_2\text{O} \), for example, a word? Or combinations of letters such as \((\text{d}l/\text{c})(\text{mc}^2/\text{E})\)? \( ^\circ \text{C} \) or \( \text{km}^2 \)? There is no simple and definitive method of distinguishing between what should properly be called words and what are simply combinations of letters with more or less unclear connections with the vocabulary of the language.

In light of the fact that there is no consensus as to where the limits of the words should be drawn, it can prove difficult to say how many words there are in a given text. This is true even when, as in this instance, the discussion is confined to words as they are written on the page, written or running words. The problem becomes even more difficult when an attempt is made to classify the words. In so doing we are asking whether one written word is in fact “the same” as another. Is the word er (is) and ear (was), for example, the same word? Yes, in a certain sense. In both cases we have an inflected variant of the same word, i.e. the verb vera (to be). In another sense, these are obviously not “the same” word. One is written e-r and the other v-a-r. In such a case it is thus necessary to examine what is the intended meaning of the word word.

There are still other factors to be considered. In a combination of letters such as BSRB (the abbreviation of the Icelandic union of public employees) there are four letters, but only three different letters. A clear distinction must be made between cases where letters, words or other units are counted without regard to repetition and cases where the intent is only to count how many different units are involved.

The concepts of type and token are used to make this distinction. In the letter combination BSRB, for example, there are four tokens but only three types. The same distinction is important in examining words in a sentence.

If a short sentence such as the following is examined:

\[
\text{Ég minni ykkur á það sem þálfæðingurinn sagði í áheyrn minni: Gætði að orðunum, þálfæðingar!}
\]

it seems apparent that it is comprised of fifteen words. If a word-counting program is used to count the number of words in this sentence the result is exactly that number: fifteen. In this instance the concept of word is used in the same way as the concept of token. Such words are commonly called running words. Thus, in the sentence above there are said to be fifteen running words. In the survey of word frequency described here the total number of running words was 519,186. In comparison the Old Icelandic Saga of Burnt-Njal contains slightly fewer than 100,000 running words.

The separation of one word from another is determined by their written form and the rule most often followed is to count continuous strings of symbols between spaces or punctuation marks as individual words. The rules which were used in this survey to determine running words are discussed in detail in section 3.1.4 of the preface.

The concept of running word is a token concept. But the words are also arranged by type and, as such, in various ways, according to the manner in which they are interpreted.

In the first place, it is possible to examine only the written forms of words without regard to their meaning or inflected form. Using this approach the number of different words in the sentence above turns out to be fourteen, as one of the words, minni, occurs twice. The concept of word form (orðmynd) is used to refer to words in this sense and there are thus fourteen such word forms in this sentence. Most of the word forms in the sentence above occur only once, but one of them occurs twice.

Computers are very efficient at counting both running words and word forms in computer-readable texts and thus many of the frequency studies which have been
carried out using computers are limited to counting these two phenomena. In so doing, however, important information is overlooked. If the former sentence is examined more closely we see, for example, that it contains two variant forms of the word málfræðingur. One of them, málfræðingurinn, occurs in the nominative case singular with suffixed definite article, the other, málfræðingar, in the nominative plural without the article. There is considerable justification for classifying these words as one and the same. It also becomes apparent that classification by word form groups together two word forms which do not represent the same word in the dictionary sense. This is the word form minni. In the former instance, “Ég minni”, it occurs as a verb but as a pronoun in the latter instance, “áheyrn minni”. Mechanical classifying of words makes no distinction here, as it is based solely on the written forms of words. The concept of lemma (fleittlorð) is used here to refer to words which differ in a grammatical sense. This is thus another type concept for words which differs from the concept of word form (orðmynd) where classification was based solely on the written form of the word. The analysis of lemmas is based on grammatical classification of the words. Words of different word classes are thus classified as separate lemmas. In the case of a noun the gender is also considered. The noun höð occurs in both feminine and neuter forms and thus counts as two different lemmas, despite the fact that the outward appearance of the word is the same, at least in some of the inflected forms.

The classification of running words into lemmas is considerably more complicated than is their classification into word forms. Computers can classify running words into word forms without difficulty, as this classification is based solely on a comparison of written forms. Classification into lemmas must be based on considerable knowledge of grammar and computers do not possess such knowledge as a matter of course. The survey which is described in this work is a pioneering effort in Icelandic frequency studies, as it represents the first instance in which the vocabulary of the study is classified into lemmas. This analysis was done partly by mechanical means and partly manually, as will be described later.

Finally, the concept of analysed word form (greiningarmynd), which is used fairly extensively in this study, should be defined. Analysed word form is used to refer to a word form together with an “analytical tag”. The main chapter of the book (pp. 3-554) contains all the word forms classified by lemma. These word forms are, in addition, classified separately if they have different inflected forms, despite the fact that they appear in the same written form. The word form haust, for instance, is entered three times under the lemma haust (p. 182). In the first example the word occurs in the nominative singular, in the second in the accusative singular and finally in the accusative plural. All these inflected forms are represented by the same word form, haust. If the intention of the study had been merely to distinguish lemmas all of these inflected forms could have been grouped together under the lemma. But as it was our aim to distinguish between all the inflected forms of words they are kept separate. We use the concept of analysed word form to refer to a word form together with its analytical tag.

To return once more to the sentence

Ég minni ykkur á það sem málfræðingurinn sagði f áheyrn minni: Gætði að orðunum, málfræðingar!

it turns out that the total number of running words it contains is 15, the number of word forms is 14, of lemmas is 14 and of analysed word forms 15.
3 The frequency survey by the Institute of Lexicography

As mentioned previously, some time has passed since the word frequency survey of the Institute of Lexicography began. In this chapter the manner in which the actual survey was conducted and the material used is described in more detail.

3.1 Conduct of the survey

3.1.1 Choice of texts

Included in the survey are one hundred texts, all similar in size, each approximately 5,000 running words. In the vast majority of cases the texts are part of a larger work.

The texts were chosen from written works which were published during the decade from 1980–1989. If an edition other than the first edition of a work was used care was taken to ensure that the first edition of the text was also published during this period. In the case of a translated work the reference year is the year in which the translation was published rather than the original.

In the beginning it was decided that an equal number of texts would be selected from five classifications, twenty from each. The classifications were as follows:

1. Icelandic fiction.
2. Translated fiction.
4. Scholarly works.

The texts selected from the fourth group were divided equally between the humanities (ten texts) and science and technology (ten texts). The texts from the fifth group were divided equally between original works (ten texts) and works in translation (ten texts).

A further requirement was that the author(s) or translator(s) of a text should neither have written nor translated any other text used in the survey.

Apart from this, the choice of text was ultimately determined by the ease of accessibility of the works, in particular whether they were available in computer-readable form. Most of the texts were obtained from the computerized text corpus of the Institute of Lexicography but several were keyboarded especially for this project.

Each selection begins with continuous written text and not, for example, with titles or chapter numbers. When the text chosen was part of a larger written work, as was most often the case, it generally, although not always, commenced where the main written text began and any pictorial or other miscellaneous material was omitted. The text always concludes with a complete sentence, most often at the end of a chapter or other clearly demarcated section of the work.

3.1.2 Material

The following is a list of the works from which samples of text were obtained. The page numbers are given if the sample is taken from a text which forms part of a larger work
and they indicate the location of the text in that work, but the exact location of each sample in the text is not given.

1. Icelandic fiction


2. Translated fiction


3. Biography and memoirs


4. Scholarly works


5. Books for children and youths

3.1.3 Alterations to the texts
The texts were carefully proofread and obvious printing errors corrected. Unusual written forms, however, which are the result of the eccentricities of the author, were not changed except as described here below.
Before the texts were analysed the following alterations were made:

1. All punctuation, except word-level punctuation, was removed. Word-level punctuation includes punctuation marks which are part of words, as, for example, the hyphen in Austur-Greñaland or BASIC-forrit.
2. Word-initial capital letters at the beginning of sentences were converted to the respective lower case ones except in cases where it is customary to write the word with an initial capital, as is the case, for example, with proper nouns. Capitals were also converted to lower-case letters in words written in capitals for emphasis, such as titles.
3. The spelling of words containing the letter ‘z’ was altered in accordance with present spelling regulations, in other words ‘z’ was changed to ‘s’ except where ‘z’ is permitted according to the regulations. In four instances a ‘t’ was dropped concurrently: breytz → breyst (twice), bøetz → bøest and sætz → sæst.
3.1.4 Distinguishing running words

The concept of running words (lesmalsörð) includes all sequences of alphanumeric characters separated by spaces or punctuation marks. There was seldom any difficulty in distinguishing one running word from another, although questions arose with regard to numbers and various symbols. The first chapter of the book (pp. 3–554) shows the manner in which running words have been distinguished. The following points may be mentioned:

1. "Normal" words refer to those which are obviously separate running words, such as og, maður, telja etc.

2. Numerals are also considered running words and the longest sequence of characters possible is included in each numeral. Plus and minus signs and percent symbols, for example, are included with the running word. 10, 10, and 10% are thus different running words.

3. In keeping with the principle of including as long a string of characters as possible in each numeral, the expression 0.7 x 10^6 is considered one running word; the same applies to 10–15, 1900–1930 etc.

4. Combinations of numerical, alphabetical and other characters which together form a single entity are considered single running words: Al_2O_3, H_2O, (d/2c)(mc^2/E)^2, km^2, Mg(OH)_2.

5. Perhaps the main point which may be questioned is the classification of mathematical equations as single running words: δ=+94.9%. The survey includes relatively few such equations, however.

6. Abbreviations are classified as one or more running words depending upon how they are read. Abbreviations of units of measurement are, however, always classified as only one running word. To give a few examples:

   BSRB is pronounced letter by letter and classified as one running word BSRB.

   H.Í is read as "Háskóli íslands (University of Iceland)" and classified as two running words: H. and Í. They are subsequently added to the lemmas háskóli (university) and Island (Iceland).

   The abbreviation e.t.u. is read as "ef til vill" and classified as three running words: e., t. and u. They are subsequently added to the lemmas ef, til and vilja.

   The abbreviation etu is read as "ef til vill" and classified as three running words: e, t and u.

   The abbreviation for the unit of measurement km/s is read as "kilómetrar á sekundu" but classified in spite of this as only one running word: km/s.

7. The manner in which words were written in the texts also determined to some extent their classification as separate running words.
3.1.5 Word analysis

Before the actual grammatical analysis began the text was divided into running words and each word allotted a separate line. Analysis was carried out by assigning to each word an analytical tag and indicating the lemma to which it belonged.

The analytical tag of a running word is a string of alphabetical characters, occasionally accompanied by a numerical character, which identify the word class, inflection and possibly other classifying characteristics of the word (cf. section 5.1).

The following example is taken from the beginning of one text:

Ég gleymi örugglega aldrei árinu 1980. þá gerðist allt.

The following is the result of the analysis:

| f p l e n | ãg | ãg |
| s g f n e l | gleymi | gleyma |
| a - a | örugglega | örugglega |
| a - a | aldrei | aldrei |
| n h e ð g | árinu | ár |
| t | 1980 | 1980 |
| a - a | þá | þá |
| s m f þ e 3 | gerðist | gera |
| f o h e n | allt | allur |

First comes the analytical tag, followed by the running word and finally the lemma. There is an empty line between sentences.

3.1.5.1 Mechanical analysis

Right from the start of the survey it was apparent that it was the analysis of words which would require the most time and effort. For this reason we decided to apply mechanical analysis with the help of computer technology as extensively as was considered feasible.

The use of mechanical analysis, although it fails to produce the correct results in many instances, is time-saving because it is easier to correct the errors manually than to carry out the entire analysis manually.

The mechanical analysis of the first fifty texts was based on the analysis used in a previous frequency study (Friðrik Magnússon 1988) which included approximately 54,000 running words. For the final fifty texts, however, the mechanical analysis was based on a corrected analysis of the first fifty which included 250,000 running words.

In addition to making use of the results of previous analyses, the mechanical analysis was based on the presumption that the inflected form of a running word can often be determined from the word form although this is seldom unambiguous. In some cases the latter part of a word suffices to determine its word class and inflected form even though the former part of the word is not known. Information on inflections can be put to good use in mechanical analysis of words which have not appeared previously.

Numerous other simple rules regarding the relationship of words in a sentence, such as rules regarding case determination and correspondence in inflection, were used in the mechanical analysis. Such rules are also of assistance in analysing previously unknown words.
The final results of the mechanical analysis are based on computations of probability, as the analysing program calculates which rules have been complied with and which have not and this determines the result.

There is considerable difficulty in assessing how exact the mechanical analysis is, in particular due to the fact that certain points of analysis were reviewed in light of previous experience. It is, however, safe to say that more than 80% of the running words are analysed completely correctly by the mechanical analysis, with regard to both their tags and lemmas. (A more exact discussion of the mechanical analysis can be found in Stefn Briem 1990.)

3.1.5.2 Correction and appraisal of the analysis

Following the mechanical analysis of each text the results were examined and corrected and words which the mechanical analysis had left unclassified were analysed. At this stage a number of doubtful questions which had arisen in the course of the analysis were resolved, the further details of which are explained in the following chapter on grammatical analysis.

4 Grammatical analysis

In order to be able to investigate the frequency of grammatical features and lemmas a detailed analysis of grammar is required. In this chapter the grammatical analysis which produced most of the frequency numbers presented in this book is discussed. The grammatical features recorded for each running word are tabulated and the problems and questions of doubt which arose in the course of the analysis are discussed. We begin with a discussion of the word class analysis in section 4.1, followed by the grammatical aspects of individual words in section 4.2 and finally the lemmas in section 4.3.

4.1 Word class analysis

This survey distinguishes between eight word classes: nouns, adjectives, pronouns, article, numerals, verbs, adverbs and conjunctions. Words which do not fall into any of these classes are grouped into two categories, foreign words and unanalysed words. The main deviation from the traditional Icelandic analysis of word classes (i.e. as this is taught in Icelandic schools) is that pronouns, exclamations and signs of infinitive are not classified as separate categories. This will be discussed in more detail later. Various cases of doubt arose in the analysis of word classes, most of which are well known from textbooks on word class analysis and it is unnecessary to engage in an extensive discussion of them here. Nevertheless, the main points of uncertainty should be mentioned, as well as the main rules which were followed in the word class analysis.

Despite the fact that various grammatical reasons will be given for the analysis used here it must be kept in mind that a number of choices were made for practical reasons to facilitate the analysis. Where two options are available for the analysis (e.g. in the determination of word class) it is tempting to choose the course that is simpler and requires less effort.

4.1.1 Prepositions and adverbs

In this survey prepositions are not treated as a separate word class but instead are classified together with the adverbs. In traditional word class analysis it is customary
to distinguish between prepositions, which govern a certain case, and adverbs which generally do not govern case. These word classes overlap considerably, however, as is pointed out by Bjarður Guðfinnsson in his discussion of Icelandic grammar (1939:99): “Most prepositions can serve as adverbs and a great number of adverbs become prepositions. This is determined by the position and the usage of the words. Certain rules determine when a particle is a preposition and when it is an adverb. It is only considered a preposition when it governs the case of its object.” A particle, however, is not necessarily a preposition even though it governs case: “Some adverbs can also govern case, without acting as prepositions” (ibid). On the other hand, prepositions “become adverbs, when their objects are removed” (ibid). In other words, prepositions always become adverbs when they do not govern case, while adverbs only sometimes become prepositions when they govern case.

Fourty-six of the fourty-seven words which Bjarður Guðfinnsson (1939:99) lists as prepositions occurred in this frequency study. Thirty-seven of these fourty-six words can also be adverbs when they do not govern the case of their object, as only nine of them always govern case. They are án, gagnvart, gegnt, gegnum, handa, móti, sakir, sökum and umfram. All of these words are relatively rare and together comprise less than 1% of the total number of instances where the fourty-six prepositions occur as running words. All of the most common prepositions can also be adverbs.

Because of the extensive overlap between prepositions and adverbs, we felt it was right to re-examine the classification of prepositions as a word class, especially with regard to other word classes. In his textbook on grammar Bjarður Guðfinnsson (1939:125–6) names six word classes, apart from prepositions, which can govern case: transitive verbs, nouns, pronouns, adjectives, numerals and adverbs. None of the words in these word classes (with the exception of some of the adverbs) fall into another word class when they govern case. In other words, case determination never decides to which word class a word belongs except in the case of prepositions and adverbs. Verbs continue to be verbs whether or not they govern case. Setja is the same word and belongs to the same word class whether or not it governs case. Accordingly, we maintain that til is the same word and belongs to the same word class whether or not it governs case. There is thus no distinction made here between adverbs and prepositions in the word class analysis and instead the words which are customarily classified as prepositions are grouped with the adverbs. In so doing we also obtain a better overview of the frequency with which these words govern case or fail to do so.

4.1.2 Exclamations

Exclamations are not considered a separate word class here, as is customary in traditional word class analysis, but are grouped with adverbs, as some adverbs can at times serve as exclamations. The exclamation æ, for instance, in the sentence Æ, þessi fjöandans höfuðverkur, is a normal adverb in the sentence Vegna minnkandi eiginfljómögnunar hafa því ordið æ háðari lánssjármögnun.

4.1.3 Sign of infinitive

The so-called sign of infinitive is classed here as a conjunction, as it is difficult to justify the distinction which has been made between it and the relative conjunction að in traditional word classification.
4.1.4 Relative pronouns

The relative pronouns sem and er, which are classified as pronouns in traditional word class analysis, are classified here as conjunctions in accordance with their position and usage. For reasons in support of this classification see Höskuldur Prætnsson (1980).

4.1.5 Adjectives and adverbs

Prepositions are not the only word class which overlaps with adverbs. Neuter singular adjectives in the nominative case (or the accusative, which is always the same as the nominative,) are considered as adverbs under certain conditions and generally it is the position and usage of the word which is used to determine its word class. Adjectives occur in a definite case and accompany a case-inflected word, while adverbs generally modify verbs, adjectives or other adverbs and have no case inflections. As an example, the neuter form fast is classified as an adverb in Markúsina hét fast við hugmynd sina, but as an adjective in hann hafti komb þut í fast horf. This classification is followed here, as it is possible to form comparatives and superlatives of some of the words when they are used as adverbs, in which case they acquire the comparative inflections of adverbs and not adjectives: hann gripur fastar um pokann. Other examples of adverbs of this sort are: hann rekki pipa allmikló, eldnöggi greip hán þad upp, langt á undan, barst þut fjöll fiskifrættir, dálitið fleiri, miklið veik.

On the other hand, adjectives in the neuter singular in the dative case are not classified as adverbs even though they are used as adverbs, e.g. lóngu sibar, miklu skemmtilegra, nógu vel, skómmu döur.

4.1.6 Past participles and adjectives

It is often difficult to distinguish between past participles of verbs and adjectives. Here we have followed the general rule of classifying such word forms as past participles of verbs unless it is obvious that they are adjectives, i.e. when the word has both the position of and is used as an adjective. To take an example, þveginn is classified as an adjective in Ég var í hvítum frakka og þveginni skyrtu, but as the past participle of the verb þvo in Æskilegt er áð likaminn sé þveginn vel og ræktileg að kvöldi til.

4.1.7 Present participles

It is often a matter of question how the so-called present participle should be treated. It is classified as a verb (or even a noun) in traditional word class analysis, yet seldom has the position of, or is used as, a verb but usually that of an adjective or even an adverb. Here we have followed the general rule of classifying a present participle as an adjective when it has both the position of and is used as an adjective: fjúgandi hátka, hækandi kaupgjáld, kaefandi skittapest, augun urðu stór og spyrjandi, hann vetlíst i grásinu hlæjandi; as an adverb when it has the position and usage of an adverb: alveg standandi hissa, þetta er vonandi tímabundni ástand, sagði konan afsakandi, Bennet swarði þut neittandi; and otherwise as a verb: verða miklu ráðandi, for lístrænn tjáningarþöf hans svaxandi.

4.1.8 Ordinals

Here ordinals are classified with adjectives and not numbers as is the case in traditional word class analysis. This is first and foremost due to the fact that it has not proved possible to distinguish between the adjective fyrr (former/earlier), fyrti (first/earliest) and the ordinal fyrstur (first) as is done in traditional word class analysis. In the Icelandic
dictionary (Íslensk orðabók handa skólad og almenningi 1983) the adjective fyrrri is said to mean “sá af tveim sem á undan fer í róð í tíma eða rúmi (the one of two which precedes in time or space)” (p. 256). This definition, however, hardly applies to the superlative form fyrrstur, for instance, in the examples of usage which are given: “í fyrrsta lagi fyrist; ekki fyrr en; með fyrrsta = í fyrrstu í upphafí”(ibid). For this reason we have followed the course of classifying the superlative fyrrstur, fyrrsti as a separate word apart from the comparative fyrrri, and classify all ordinals as adjectives. This simplifies the analysis of these words appreciably and saves considerable work in editing the mechanical analysis.

4.1.9 Pronouns and adverbs

As was previously mentioned, adjectival word forms in the neuter are classified as adverbs when they are used as adverbs and modify verbs, adjectives or other adverbs, as it is at times possible to form comparatives and superlatives of these words just as of normal adverbs. This is not the case with pronouns (especially indefinite pronouns) which can also be used as adverbs and modify verbs, adjectives or other adverbs, as they cannot be compared as can adverbs. These word forms are thus classified as pronouns. The following are several examples of this usage of the pronouns: allnokkuð kvönhollur, allt frá landnámssöld, eittvæðu fáll, Vaskur skammatist sin augsýntilega ekkert, honum var nokkuð brugðóð, nokkrum síðar, alla heldur, allra best, alls ekki, einna dýpst. There are, nevertheless, examples where pronominal word forms are classified as adverbs, e.g. alls when it means ‘a total of’ or ‘all together’ (samtals), annars when it means ‘otherwise’, ‘in fact/rather’ and því when it means ‘thus/for this reason’, e.g. in alls 240 kr., annars hefði hún ekki verið þarna, ég bis þíku því að vera viðbúinir.

4.1.10 Einn (One)

The word einn has a widely varying and often rather unclear meaning. Here we distinguish between its usage as a numeral, indefinite pronoun and adjective. Einn is most often classified as a number, e.g. in eitt litið ollumálverk, einn af máttarstólpum þessa litla samfélags, einn eða fleiri, einn og einn háður, as an indefinite pronoun when it means ‘an’ or ‘one’, for example in eitt helsta viðfangsfini hans, svo fátt eitt sé nefnt, einu sinni þekkti ég strák, einna dýpst, eins konar, dag einn, and otherwise as an adjective when it means ‘one, the same, alone, only’, etc., as in: allt í einu, það einu sem hún var í, hégömt einn, ýmist einn eða með ðörum, þarna býr hann einn í nýlegu einbylislúi. It must be admitted, however, that it is often difficult to distinguish between the various usages of the word.

4.1.11 Adverbs, adjectives and nouns

In this section some examples of adverbs, adjectives and nouns which have proved difficult to classify in the word class analysis are discussed. These words are of various sorts and often there is no other reasoning upon which to base the decision regarding classification than subjective personal judgement. As a main rule, the words were separated into word classes according to their usage and syntactic position in the sentence, but word forms are classified as nouns if they can reasonably be connected with a nominative reference form. The word konar is, however, classed as a noun although it has not proved possible to find it any other or better reference form than konar.

1. The following words are classified as adverbs (when they are in the position of an adverb and used as such): allskostar, annarsstóðar, dalengdar,
2. The following words are classified as adjectives (when they are in the position of adjectives and used as such): allskonar, ámótta, bêlúltans, einhverskona, elskuhrartans, fennskonar, gamaldags, hverskonar, stóreflis, sundurorða, tvímastra, utangátta, þesskonar, örmagna.

3. The following are examples of word forms which are classified as neuter nouns (even though they may be in the position of an adverb or adjective): eftirléttis, fjáðaræm, kvenkyns, óskýp, ótal, síðdegis, undra; these are classified as masculine nouns: afburða, andskotans, aumingja, djöfjuls, fjandans, fjarans, konar, veslings; and these as feminine nouns: alvörð, elsku, furðu.

4.1.12 Foreign words

Foreign words which occur in a continuous Icelandic text are analysed as if they were Icelandic words (note, however, section 4.2.1 here below with regard to the classification of nouns). If they do not occur in a continuous Icelandic text, or occur as inserts in the text, they are analysed as foreign words. These could take the form of:

1. whole sentences: Home is where the heart is, Et sygt samfundssystem skaber i første omgang syge oprørere, Omnis mundi creatura quasi liber et pictura nobis est in speculum,

2. exclamations and addresses: love, adieu, salut, cara, au revoir,

3. long titles and names: Analytische Sozialpsychologie und Gesellschaftstheorie, Un hiver à Majorque, oever farmasöytiske spesialpreparerer, De rerum natura, Groundwater systems in Iceland traced by deuterium,

4. direct speech in a foreign language: “D'accord. Je sais. Mais nous,” stamadi Houston,

5. foreign proper nouns which are written with an initial lower-case letter, such as de, da, dal, aga, di and von in the following names: Pierre de St. Laurent, Joakim Soares da Cunha, Castel dal Monte, Abdí aga, Accademia di Belle Arti, Joachim von Ribbertrop.

4.1.13 Unclassified words

Those words which can neither be classified according to word classes nor rightly considered to be foreign in nature are left unanalysed. This applies primarily to individual alphabetical characters such as a, b, c in lists, abbreviations which it has not proved possible to interpret, chemical formulas such as Al2O3H5 and symbols of various sorts, such as =, :, → and °.

4.2 Grammatical information given for individual word classes

In this section we discuss the grammatical information which is given for each running word and the difficulties which have arisen with regard to analysis. The main principle followed in the analysis was to attempt to fully analyse each and every running word. As might be expected, this has not been entirely possible, and words which it proved
impossible to analyse ended up either being classified as foreign words or left unana-
ysed (see above). Accompanying each word class is a table showing the symbols which
were used in the grammatical analysis. A complete table of all the word classes and
analytical aspects is given in section 5.1.1.

4.2.1 Nouns

<table>
<thead>
<tr>
<th>Column</th>
<th>Category</th>
<th>Analytical symbol — information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Word class</td>
<td>N-noun</td>
</tr>
<tr>
<td>2</td>
<td>Gender</td>
<td>K-masculine, V-feminine, H-neuter, X-gender unspecified</td>
</tr>
<tr>
<td>3</td>
<td>Number</td>
<td>B-singular, F-plural</td>
</tr>
<tr>
<td>4</td>
<td>Case</td>
<td>N-nominative, O-accusative, D-dative, G-genitive</td>
</tr>
<tr>
<td>5</td>
<td>Article</td>
<td>G-with suffixed definite article</td>
</tr>
<tr>
<td>6</td>
<td>Proper nouns</td>
<td>M-name of person, D-place name, S-other proper noun</td>
</tr>
</tbody>
</table>

The gender, number and case of each noun is given, whether or not it has a suffixed
definite article or is the name of a person, place name or other type of proper noun.

There are three genders: masculine, feminine and neuter, and wherever it has not proved possible to determine the gender of a word, despite the fact that it occurs
within the context of a sentence and in a certain case and number, it is classified as
gender unspecified but analysed in other aspects like a normal noun. These nouns are
of three types: 1) Foreign place names and other proper nouns, and even several
foreign personal names where the gender of the individual is not evident. The fol-
lowing are examples of such words: Vesterbrogaade, Boston, Politiken, Telegraph, CIA,
Karamzin, Friedman. 2) Foreign loanwords such as: rugby, asparagus, slapstick, sauna.
3) Abbreviations of chemical elements such as: Li, Na, S, Ar.

It should also be mentioned that Icelandic family names and foreign personal and fam-
ily names are analysed according to the natural gender of the individuals concerned.
Thus, Parker and Nordmann are classified as masculine in Parker var samarrekhinn
maður and til Jóns Nordmann plantóleikara, but as feminine in sagði frá Parker and
Jórun Nordmann og börn hennar.

In addition, abbreviations such as ASÍ, BHM, MA are classified according to the
gender of the principal noun. ASÍ, BHM, KFUM are thus classified as neuters since the principal
nouns in these combinations are neuter alþjóðasamband, bandalag, félag, but MA, MR,
MS are classified as masculine as mennaskóli is the principal noun of these noun
combinations.

No noteworthy difficulties arose in the classification of nouns as singular or plural and
the suffixed definite article is even easier to deal with. Ascertaining the case is naturally
a more complicated matter, but in this regard it was primarily individual examples
which caused difficulty. These generally occur in set phrases and expressions where
it is difficult to add a suffixed definite article to the noun or use an adjective or other
determining word which could reveal the case better than the noun, such as í pokkabót,
and also where nouns are used to lend emphasis, as in óskip mjó, aumingja konunum
dauðbróð, álþóru listamáður.

Nouns are also classified as to whether they are common nouns, names of persons or
places, or other proper nouns (i.e. proper nouns which are neither names of persons
nor places). The last three categories are referred to collectively as proper nouns to
distinguish them from the first, the common nouns. The fairly simple rule of classifying
as such only those proper nouns which are written with an initial capital was used
to distinguish proper from common nouns. By so doing it proved fairly simple to distinguish proper from common nouns, except in the very few instances where the words occurred at the beginning of a sentence and were therefore written with an initial capital without this being necessarily an indication as to whether the word was a proper or common noun.

As a result of this rule only nouns were classified as proper nouns and also, where a noun or name was comprised of several nouns, the first of the series is classified as a proper noun but not the others unless they are written with an initial capital. Thus in the following compound nouns only the first word is classified as a proper noun: Gráskinn frá hin meiri, Prepin prettanin, Bolstur frá bernskutó, but all the nouns which have an initial capital in: Ferðabók Eggerts og Bjarnas, Bejarútgarð Hafnarfjarðar, Alþýðusamband Íslands. The same applies to foreign names which are comprised of several running words with the exception of the fact that in these cases all the words written with initial capitals are classified as proper nouns whether or not they are considered as nouns in the foreign language. Thus, all the words written with an initial capital in the following names are classified as nouns and proper nouns: Accademia di Belle Arti, British Museum, Gare de L'Ouest, Jan Mayen, Joakim Soares da Cunha, Los Angeles, New York, The Daily Telegraph, Walter von Knebel.

Proper nouns are then classified into the subcategories names of persons, place names and other proper nouns, as was previously mentioned:

1. Names of persons include names of people, animals and other living beings, such as: Ádalbjörg, Egill, Fási, Janet, Skalla-Grimur, Vaskur.

2. Place names include the names of countries, cities, mountains, rivers, streets, houses, etc., such as: Akranes, Bókhlöðustigur, Cambridge, Eldborgarhraun, Engey, Faxafloi, Faxa, Laekjartorg, Neptunus, Norður-Atlantshaf, Portugal, Skardsfjörvarit, Skólavördubólti, USA, Vesterbrogaade, Zimbabwe.

3. Those proper nouns which can neither be considered as names of persons nor places are classified as other proper nouns. Examples of such words are: Afsatrýggingsstíður, Austföngur, Blöndahlskaramella, Bretavínna, Buick, Dani, Eimskipafélag, Framsöknarflokkr, Gestapó, Háðlarforleiður, Jónsmessa, Kötlugos, Lugerbyssa, Norðurlandarð, Sogamýrarvgur, Sturlungar, Öskubuskusaga.

4.2.2 Adjectives

<table>
<thead>
<tr>
<th>Column</th>
<th>Category</th>
<th>Analytical symbol - Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Word class</td>
<td>I - adjective</td>
</tr>
<tr>
<td>2</td>
<td>Degree</td>
<td>P-positive, M-comparative, E-superlative</td>
</tr>
<tr>
<td>3</td>
<td>Declension</td>
<td>S-strong, V-weak, O-indeclinable</td>
</tr>
<tr>
<td>4</td>
<td>Gender</td>
<td>K-masculine, V-feminine, H-neuter</td>
</tr>
<tr>
<td>5</td>
<td>Number</td>
<td>E-singular, F-plural</td>
</tr>
<tr>
<td>6</td>
<td>Case</td>
<td>N-nominative, O-accusative, P-dative, R-genitive</td>
</tr>
</tbody>
</table>

The degree and declension of each adjective is given, in addition to its gender, number and case. Adjectives have three degrees: positive, comparative and superlative. Their declension is either weak, (hvit kaninarn) or strong (hvit kantina), or the adjective is indeclinable, i.e. its form does not change in declining. Examples of indeclinable adjectives are: agndofa, farlana, forvída, framandi, frávida, gjaldræta, hissa, hringlaga, lotningarvekjanandi, ómálga, sænskumelandi, yfirfrýrandi, yxna. No difficulties of note arose in the classification of these inflectional aspects of adjectives.
4.2.3 Pronouns

<table>
<thead>
<tr>
<th>Column</th>
<th>Category</th>
<th>Analytical symbol - Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Word class</td>
<td>F-pronoun</td>
</tr>
<tr>
<td>2</td>
<td>Subcategory</td>
<td>A-demonstrative pronoun, B-indefinite demonstrative pronoun, E-possessive pronoun, O-indefinite pronoun, P-personal pronoun, S-interrogative pronoun, T-relative pronoun.</td>
</tr>
<tr>
<td>3</td>
<td>Gender/Person</td>
<td>K-masculine, V-feminine, H-neuter / 1-1st pers., 2-2nd pers.</td>
</tr>
<tr>
<td>4</td>
<td>Number</td>
<td>E-Singular, F-plural</td>
</tr>
<tr>
<td>5</td>
<td>Case</td>
<td>N-nominative, O-accusative, D-dative, E-genitive</td>
</tr>
</tbody>
</table>

Pronouns are divided into seven subcategories: personal pronouns, possessive pronouns, demonstrative pronouns, interrogative pronouns, indefinite pronouns, indefinite demonstrative pronouns, and relative pronouns. This classification departs from the traditional classification of pronouns in three respects: Firstly, the reflexive pronoun sig is not considered as a separate subcategory but included instead with the personal pronouns (as a reflexive personal pronoun) in accordance with the reflexive possessive pronoun sinn, which has always been included with the possessive pronouns. Secondly, the pronouns hvílikur, samur, sjálfur, slikur and þevlikur are classed as a separate subcategory, indefinite demonstrative pronouns, whereas these words have generally been classed as some sort of subcategory of demonstrative pronouns (with the possible exception of hvílikur). Thirdly, there is only one relative pronoun, hvör, as sem and er are classed as conjunctions as has previously been explained.

Apart from this pronouns are classified as follows: for each personal pronoun in the 1st and 2nd person the person, number and case is given; for other pronouns the gender, number and case are given.

The principal difficulties in analysing pronouns arise when they are classified into subcategories; there are a fair number of examples where the same word form can represent different pronouns. This is most apparent in the case of the third person pronoun in the neuter singular and in all its plural forms, as in the demonstrative pronoun só, i.e. the word forms það, því, þess, þeir, þá, þeim, þeirra, þær, þau. The difference between these pronouns is generally evidenced by the fact that the personal pronoun most often stands alone while the demonstrative occurs with another case-inflected word. It is difficult to form definite rules describing the difference between the pronouns in such cases, although it can be said that in most cases it is the demonstrative pronoun and not the personal pronoun which introduces a subordinate clause. Það is, for example, a demonstrative pronoun in: Eg nenni ekki að vera að bera á bóð það sem enginn vill bóða.

4.2.4 Independent article

<table>
<thead>
<tr>
<th>Column</th>
<th>Category</th>
<th>Analytical symbol - Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Word class</td>
<td>G-article</td>
</tr>
<tr>
<td>2</td>
<td>Gender</td>
<td>K-masculine, V-feminine, H-neuter</td>
</tr>
<tr>
<td>3</td>
<td>Number</td>
<td>E-singular, F-plural</td>
</tr>
<tr>
<td>4</td>
<td>Case</td>
<td>N-nominative, O-accusative, D-dative, E-genitive</td>
</tr>
</tbody>
</table>

There is only one independent article in Icelandic, hinn, which is analysed according to gender, number and case. Here it is considered a separate word class, as in traditional word class analysis, but it could well have been classified as a subcategory of the pronouns due to its position and usage within the noun phrase.
4.2.5 Numerals

<table>
<thead>
<tr>
<th>Column</th>
<th>Category</th>
<th>Analytical symbol - Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Word class</td>
<td>1-numeral</td>
</tr>
<tr>
<td>2</td>
<td>Category</td>
<td>1-cardinal</td>
</tr>
<tr>
<td>3</td>
<td>Gender</td>
<td>K-masculine, V-feminine, H-neuter</td>
</tr>
<tr>
<td>4</td>
<td>Number</td>
<td>R-singular, P-plural</td>
</tr>
<tr>
<td>5</td>
<td>Case</td>
<td>N-nominative, O-accusative, D-dative, E-genitive</td>
</tr>
</tbody>
</table>

Numerals are divided into two groups in this study, cardinal numbers and other numerals (as was previously mentioned ordinal numbers are classified as adjectives). There are three main types of other numerals:

1. Dates, numbers and other indeclinable numerals such as árið 1843, klukkan þrjú, númer prettán, Ránargata 18, Stóð 5. These numerals are not declined even though the accompanying word may be declined: árið 1843, árinu 1843, ársins 1843. The numeral remains unchanged and always ends with þrjú although the noun is declined. These numerals are thus neither declined according to gender or case although they appear in the nominative/accusative neuter form of the respective cardinal number. Other indeclinable numbers which can be mentioned are chapter numbers, degrees of various sorts (60°C, 273°K, 64°30', −69°202) and various numbers which do not fall under any other category (1-7-2-7).

2. Percentages such as 10%, 1.8%. These numerals can admittedly be declined, but since it is uncertain how the percentage sign should be read, as prósent or af hundrað, they are not further analysed.

3. Numbers which precede other numerals such as hundrað and þúsund, in: tju þúsund hermann, fjögur hundrað milljónir líra, tvö hundrað hrjáðir einstaklingar. In these examples the numerals hundrað and þúsund are classified as numerals representing the same gender, number and case as the nouns which they accompany (hermann, milljónir, einstaklingar). The numerals, however, are not declined in accord with the nouns: tvö hundrað hrjáðir einstaklingar (nominative), um tvö hundrað hrjáða einstaklinga (accusative), frá tvö hundrað hrjáðum einstaklingum (dative), til tvö hundrað hrjádra einstaklinga (genitive).

These three types of numerals are not further analysed (and are thus marked only with the word class t), while the cardinal numbers are classified according to gender, number and case. The division between singular and plural of cardinal numbers differs from that of other inflected words (with the exception, perhaps, of a very few nouns) in that they are not declined in singular and plural; instead einn (one) and compound numbers which end in einn (one), such as tuttuguogeinn (twenty-one), 528.431 are singular, while other cardinal numbers are plural.
4.2.6 Verbs

<table>
<thead>
<tr>
<th>Column</th>
<th>Category</th>
<th>Analytical symbol - Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Word class</td>
<td>S-verb (except for past participles)</td>
</tr>
<tr>
<td>2</td>
<td>Voice</td>
<td>G-active, M-middle</td>
</tr>
<tr>
<td>3</td>
<td>Mood</td>
<td>N-indefinite, B-imperative, F-indicative, V-subjunctive, S-supine,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L-present participle</td>
</tr>
<tr>
<td>4</td>
<td>Tense</td>
<td>N-present, P-past</td>
</tr>
<tr>
<td>5</td>
<td>Number</td>
<td>E-singular, F-plural</td>
</tr>
<tr>
<td>6</td>
<td>Person</td>
<td>1-1st person, 2-2nd person, 3-3rd person</td>
</tr>
</tbody>
</table>

The voice and mood of each verb is specified. There are two voices, active and middle which require no further explanation (see, however, number 10 of section 4.3 with regard to reference forms and lemmas below).

There are seven moods: nominative, imperative, indicative, subjunctive, supine, past participle and present participle. The difficulty in distinguishing between participles and adjectives was mentioned previously and need not be repeated here, but the difference between the supine and past participle should be explained.

A distinction is made between an indeclinable supine and a declinable past participle. The verbs *hafa (to have)* and *geta (to be able)* are followed by a supine, as in the sentences *pau höfðu ekkert rœtt um sumarfrœð*, *húnn gat ekki annað en brosæð*, but verbs such as *vera (to be)* and *verða (to become)* are followed by past participles (e.g. in the passive voice), as in the sentences *hann var kominn með hærraklipingu, svo var hvískrini haldîð af*.

In addition to their moods and voices, verbs in a "personal mood" (in the indicative, subjunctive and imperative) are analysed according to tense, number and person, the past participle of verbs are analysed according to gender, number and case, and verbs in other moods (the infinitive, supine and present participle) are not further analysed (except that the past tense infinitives of the verbs *munu* and *vilja* are indicated).

4.2.7 Adverbs

<table>
<thead>
<tr>
<th>Column</th>
<th>Category</th>
<th>Analytical symbol - Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Word class</td>
<td>A-adverb</td>
</tr>
<tr>
<td>2</td>
<td>Degree</td>
<td>M-comparative, E-superlative</td>
</tr>
<tr>
<td>3</td>
<td>Class/case governor</td>
<td>A-does not govern case, U-exclamation / O-governs accusative,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P-governs dative, E-governs genitive</td>
</tr>
</tbody>
</table>

The extensive overlap between adverbs and other word classes has previously been described and the main difficulties in determining the limits of this word class mentioned. When the word class analysis is complete there are few problems regarding the analysis of adverbs.
To begin with, the comparatives and superlatives are specially marked. Adverbs are subsequently analysed according to category or by the cases they govern into five groups: 1) those which do not govern case, 2) those which govern the accusative, 3) those which govern the dative, 4) those which govern the genitive and 5) exclamations. Several points of doubt do admittedly arise where it is not always clear whether the words govern case or not due to the fact that the noun phrases have been separated from the case governor or simply omitted. In dependent clauses, for example, it is common for a nominal, the case of which is governed by an adverb, to be omitted: veitingahusdi sem hún hafði fyrst farið inn á __________. In this and other similar cases where it is obvious that the noun phrase has been omitted, the adverb is considered as governing case (the accusative in this instance). This is, in fact, in accordance with the traditional analysis except for the fact that there it would be termed a "preposition" and considered to be governing the case of the "relative pronoun".

### 4.2.8 Conjunctions

<table>
<thead>
<tr>
<th>Column</th>
<th>Category</th>
<th>Analytical symbol - Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Word class</td>
<td>c-conjunction</td>
</tr>
<tr>
<td>2</td>
<td>Category</td>
<td>N-sign of infinitive, T-relative conjunction</td>
</tr>
</tbody>
</table>

Conjunctions are not analysed but the relative conjunctions *sem* and *er* are specifically identified, as is the sign of the infinitive *að*.

### 4.2.9 Foreign and unanalysed words

<table>
<thead>
<tr>
<th>Column</th>
<th>Category</th>
<th>Analytical symbol - Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Category</td>
<td>K-foreign word</td>
</tr>
<tr>
<td>1</td>
<td>Category</td>
<td>X-unanalysed word</td>
</tr>
</tbody>
</table>

As has been previously discussed, the words which cannot be classified in the eight word classes are either classified as foreign words or left unanalysed. These two categories are not further analysed.

### 4.3 Reference forms and lemmas

The aim of this survey is not merely to study the frequency of word forms and grammatical aspects but the frequency of lemmas as well. In order to make this possible it is necessary to group together various word forms of the same lemma. This is done by creating a special reference form for each running word which indicates to which lemma the running word belongs. Here the concept of reference form is used only in this sense: a form which indicates to which lemma the running word belongs. The reference form for case-inflected words is in the masculine nominative singular and the reference form for verbs is in the infinitive. Thus *unglingur* is the reference form for the running word *unglingarnir, fullorðinn* the reference form for the running word *fullorðna, dverða* for the running word *dverðu*, etc.

In the following paragraph the general rules which were followed in creating reference forms are described along with the main difficulties and points of doubt which have arisen, which could be classified as one of two types: a) that "various" words can have the same lemma and b) that the "same" word can have different lemmas.
1) The difficulties in setting up reference forms are connected with the distinctions made between words within the word classes. When the word class of a running word has been determined it makes little difference if its reference form is the same as a reference form in another word class since lemmas from different word classes can never be considered as belonging to the same lemma. In other words, even though the reference form for the feminine noun saga is the same as for the verb saga, these words would not be considered as the same lemma because their word classes are not the same.

In the same manner nouns of different genders cannot be considered as belonging to the same lemma even though their reference forms are the same. Thus the feminine noun ár/árín and the neuter noun ár/árð belongs to different lemmas despite the fact that their reference forms are the same, ár. The same can be said of egg/egginn (fem.) — egg/eggð (neu.), leiði/leiðinn (masc.) — leiði/leiðð (neu.), reiði/reiðinn (masc.) — reiði/reiðð (fem.), and other similar pairs.

Apart from this the rule applies that words in the same word class, and nouns of the same gender, which have the same reference form and the same inflections are considered to belong to the same lemma even though their meanings may be different. A different meaning alone is insufficient to distinguish between lemmas. Examples of this are the neuter noun lag which is considered as belonging to the same lemma whatever its meaning, as, for example, in the following phrases: allt í lagi, einkum og sér í lagi, í mesta lagi, hvort í snu lagi, í fjasta lagi, utan laga og rétar, eins og jaxi í laginu, í eftasta lagi vokvans, að stó hann út af laginu, fallegasta lágið sem hún kunn.


Words which have the same reference form and the same meaning but different inflections are, however, not separated and are classified under the same reference form. An example of this are family names such as Gunnarsson which has the same reference form whether it is of foreign origin and not case-declined or an Icelandic patronymic which is inflected in the dative and genitive cases.

3) Nouns which have different meanings in the singular and plural are entered under the same reference form in the singular. Examples of this are: diak/dið, fai/frít, gagn/gagn, gög, lag, lág.

4) Words which are spelled in various ways (either where one spelling reflects the pronunciation and the other does not, or where two different spellings are possible without necessarily reflecting a difference in pronunciation) are combined under the same reference form and traditional spelling is adopted for the reference form. The following are examples of this, with the traditional spelling (i.e. the reference form) given first, followed by the less traditional: almennilega/almenilega, bleya/bleyja, Dyrholæy/Dyrholý, ég/égg, fjöruthú/fjörtú, fransskbrauð/fransskbrauð, giðs/gips, grafikyr/grafikjur, grenja/gønja, guð/gvû, Grímur/Grímsur.
The same applies to words which occur both with or without a hyphen; they are generally entered under the non-hyphenated reference form: bang / bang-bang-bang, dúðufugl / dúðú-fugl, Kaspián / Ka-ka-kaspián, mahóníborð / mahont-borð, nettógjald-eyrismagn / nettó-gíjálðurýsígn. This also applies to words with a slash with the result that svart/hvítur is entered under the reference form without the slash: svart/hvítur.

Exceptions to the above rule are of three kinds:

1. The masculine noun kall and the feminine noun kelling are entered under these reference forms and not under the reference forms kort and kerling. The same applies to various compounds with kall as their final node: gormakall, hörkall, járnkall, kerfiskall, Kinakall, sprellkall, tikall, tóbakskall, tröllkall, öskukall. It so happens that none of these compounds occurs with -kort as their final node.

2. A few words which deviate markedly from the traditional spelling are not entered under the “proper” reference form, as it is doubtful whether they should be considered the same words when their spelling has been altered to such an extent: ond, onefír, oní, soldób, soldíll.

3. The spelling of personal names determines their reference forms. The following pairs of names (or even three varieties) are thus examples of different lemmas: Alfred / Alfred, Annie / Anny, Bennet / Bennett, Carol / Carole, Carl / Karl, Eggerts / Eggerz, Halfdán / Halfdán / Halfdán, Hendriks / Hendrix, Janice / Janis, Joseph / Jósef / Jósef, Julia / Júlia, Lúdvig / Lúdvík, Sofía / Sofia, Solví / Solví, Valdemar / Valdimar.

There are two exceptions from this rule: the running words jesum and Jésú are both entered under the reference form Jesus; in addition the running word Zoega is entered under the reference form Zoega.

5) Compound words which are identical except for the fact that the initial node occurs either as a root form or in the genitive form (in the singular or plural), or as the root together with a hyphen, are classified as two separate reference forms. Examples are: abgerðalykill / abgerðarlykill, annarstáðar / annarstáðar, augnálok / augnálok, byggjagíðsnaður / byggjargíðsnaður, fiskinjóí / fiskinjóí, kristalglas / kristalglás, náttúrlega / náttúrleiga, rannsóknarstofa / rannsóknastaða, síðferðilegur / síðferðilegur, tilvölfjumarkennardur / tilvölfjumarkennandur.

6) There are words which have the same meaning but different inflections, with the exception of some common inflected forms where it is not possible to distinguish between them. Examples of such words are hólim og hólmur, which share to some extent the same meaning but have only partially similar inflections and do not have the same reference form. On the other hand, the plural forms of these words are identical, with the result that it is hardly possible to say that a form such as hólmum represents one word rather than the other because of their similar meanings, especially if this is the only example of the word which occurs in the text. Words of this sort are here entered under a common reference form with parentheses and/or a slash in an attempt to show both the reference forms at once. Examples of such reference forms are: dollar(i), ey(ja), éttaeta, flott(ur), hólimi-ur, kærleiki-ur,
lær(i), meiðsl(i), meir(a), mey(ja)/maer, rep(i), smið(i), smyrs(i), systkin(i), umlukinn/-lukur.

This is only done if a common inflected form occurs. If such is not the case (e.g. if only one inflected form occurs and this can only belong to one of the two reference forms) only a single reference form is entered. The following are examples of this with the word forms which occur in parentheses: dvergasmíð (dvergasmíð), hljóðferasmiði (hljóðferasmiði), kvartdollar (kvartdollars), Akurey (Akureyjar), haldreipi (haldreiði), Gunnarshólimi (Gunnarshólima). There are also cases where two or more word forms occur, none of which can belong to two reference forms. In such a case two different reference forms are entered and there are thus two lemmas. Examples of this are (with the word forms which occur in parentheses): fæleiki (fæleikinn) — fæleikur (fæleik), hellagleiki (hellagleiki) — hellagleikur (hellagleik), myndugleiki (myndugleika) — myndugleikur (myndugleik), sannleiki (sannleika, sannleika, sannleikann, sannleikanum, sannleikans) — sannleikur (sannleikur, sannleikurinn, sannleik), sjúkleiki (sjúkleika) — sjúkleikur (sjúkleikur), veruleiki (veruleika, veruleikinn, veruleika, veruleikann, veruleikanum, veruleikans) — veruleikur (veruleik).

7) Names of persons and places are entered under separate reference forms despite the fact that an identical common noun occurs in the text, or exists. For example, the personal name Álfur is entered under the reference form with an initial capital, separate from the common noun álfr. Examples of names of persons of this sort are: Bóra, Björn, Fjóla, Gestur, Lina, Prödustur. Examples of place names are: Bakki, Brekka, Nes, Tangi, Vik, Prengslí.

If the same noun can be both the name of a person and a place it is entered under the same reference form, as, for example, in Venus, Virginia, and if the corresponding common noun occurs it is entered under another reference form, separate from the names of persons or places. Examples of this are: Jökull — jökull, Ormur — ormur, Stein — steinn, Ólfrur — álfrur. Words of this sort are, however, only classified under the same reference form if they are of the same gender, cf. the discussion in 1) with regard to the fact that nouns of different genders cannot be classified under the same lemma despite the fact that their reference form is the same. Thus, the following nouns are considered as different lemmas: Berg (name of person, masculine) — Berg (place name, neuter) — berg (common noun, neuter), Jean (personal name, masculine) — jean (place name, gender undefined).

8) Those proper nouns which are neither names of persons nor places are classified here under the heading “other proper nouns” (referred to either as such or as “proper nouns” in the following discussion). They are only entered under separate reference forms if a corresponding common noun or name of person or place does not occur in the texts, otherwise they are entered under the same reference form as that common noun or name of person or place. Examples of proper nouns of this sort which are entered under separate reference forms (with an initial capital) are: Afrtryggingsjóður, Björgunarnarfjöll, Hafnrörðingur, Islandsklukka. Examples of proper nouns of this sort which are entered under the reference form of a corresponding common noun are: Andi, Fæðing, Liðrasveit, Sjúkrahús, Tilraun, Ævisaga. Examples of proper nouns of this sort which are entered under the reference form of a corresponding name of a person are: Freyja, Galileo, Iðunn, Snorri, Trausti. Examples of proper nouns of this sort which are entered under the reference form of a corresponding place name are: Hekla, Helgafell, Keflavík, Reykjanes, Vatnajökull.

If examples occur where the corresponding word can be either a proper noun, a
common noun or the name of a person or place, the proper noun is entered under the reference form of the common noun. This is the case with the proper nouns drottning and frón even though there are examples of corresponding place names, and the proper nouns Disa, Haukur and Hlif are entered under the reference forms disa, haukur and hlif even though examples of corresponding names of persons occur.

9) In section 4.2.1 it was pointed out that the gender of Icelandic family names and of foreign personal and family names is analysed according to the natural gender of the name bearer. It was also pointed out in 1) here above that nouns of different genders cannot be classified under the same lemma even though their reference forms are the same. As a result the name of a person which occurs both in masculine and feminine forms is considered two reference forms. Examples of such names are: Alice, Brown, Cardone, Eggerz, George, Hansen, Kranz, Liang, Möller, Nielsen, Nordmann, Parker, Thorarensen, Thorsteinsson, Willard.

10) No special reference forms are entered for verbs in middle voice unless the corresponding active voice does not exist. The word form komast, for example, is entered under the reference form koma and the word form minnast under the reference form minna; the word form heppnast is, however, entered under the reference form heppna and the word form farndist under the reference form farndast since the verbs *heppna and *farn do not exist.

11) Numerals occupy a special place among the word classes due to the fact that they can be represented in at least three ways: 1) in Arabic numerals (21), 2) in Roman numerals (XXI) and 3) in words which indicate how the number is to be read (tuttugu og einn). As a result, when the text is divided into individual running words, compound numerals are analysed into many running words when they are represented in words (tuttugu og fimm, for example, is classed as three running words) but not when they are represented in Arabic (e.g. 23) or Roman numerals (e.g. XXIX). As a result the frequency of numerals is dependent to a certain extent on how the compound numerals are represented.

Here the written numbers and Arabic and Roman numerals are not classified under the same reference form and thus 6, Vf and sex are three different lemmas. Unanalysed numerals (for which the only word class indication is n) are entered under the same reference form as the corresponding cardinal number if this occurs.

12) So-called compound conjunctions (til þess að, vegna þess að, svo að o.fl.) are here divided into single running words and subsequently analysed individually (as adverbs, pronouns, conjunctions, etc.) and each entered under its own reference form. At times, however, these multiple conjunctions are written as one word and are then analysed as one running word (and this word form entered as a reference form) and classified as conjunctions, e.g., einsog, parsam, tilað, hviald, aðuren.

13) At times it can prove impossible to find a reference form in the masculine singular nominative form for an adjective, even though it is not an indeclinable adjective (see section 4.2.2 regarding the grammatical aspects of adjectives). These are adjectives which have become "frozen" in individual expressions or combinations, and/or are used without reference to another case-inflected word having gender, number or case, or with reference to a complement clause (which is always in the neuter singular) or það. These adjectives only occur in the neuter singular nominative or accusative and it is this word form which is entered as reference form. Examples of this are word forms such as ann, abótað, flókur, óhætt, ómótt and vîðvar, in sentences such as: hann lét sír mjögg ann um fjoðskyldu sina, bændur eru hvattir til að lagfjara strax það sem
5 Concerning individual chapters of the book

This section contains a survey of the chapters in the main part of the book (from p. 3 onwards). It deals with the material in each chapter and discusses the manner in which it was prepared from the results of the frequency survey. The chapters are numbered 1–14 and the number of each is given in brackets.

5.1 Lemmas and analysed word forms [1]

5.1.1 Comprehensive vocabulary list of the survey

The first chapter contains the comprehensive list of the vocabulary of the survey. It runs to a total of 552 pages. All the lemmas which were analysed in the one hundred sample texts, together with the analysed word forms of each lemma which occurred in the texts, are included in the chapter. A short extract of the chapter material is given in the following:

<table>
<thead>
<tr>
<th>Lemma</th>
<th>Number</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>fóstra</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>fóstra</td>
<td>NVEN</td>
<td>1</td>
</tr>
<tr>
<td>fóstra</td>
<td>so</td>
<td>1</td>
</tr>
<tr>
<td>fóstrar</td>
<td>SGFNE3</td>
<td>1</td>
</tr>
<tr>
<td>fóstri</td>
<td>no</td>
<td>3</td>
</tr>
<tr>
<td>fóstri</td>
<td>NKEN</td>
<td>1</td>
</tr>
<tr>
<td>fóstra</td>
<td>NKEO</td>
<td>2</td>
</tr>
<tr>
<td>fóstra</td>
<td>NKEE</td>
<td>1</td>
</tr>
<tr>
<td>fóstur</td>
<td>no</td>
<td>6</td>
</tr>
<tr>
<td>fóstur</td>
<td>NHEO</td>
<td>5</td>
</tr>
<tr>
<td>fóstri</td>
<td>NHEE</td>
<td>1</td>
</tr>
<tr>
<td>fóstrus</td>
<td>NHEE</td>
<td>3</td>
</tr>
</tbody>
</table>

There are four lemmas in this extract, the feminine noun fóstra, the verb fóstra, the masculine noun fóstri and the neuter noun fóstur.

The word list is set up with the lemmas arranged in alphabetical order and the analysed word forms then arranged in a special "grammatical order". This can be seen in the extract above; there are three analysed word forms of the noun fóstri, in the nominative, accusative and genitive singular without an article. They are arranged in this order as indicated by the final letter of the analytical tag, n (nominative), o (accusative) and e (genitive). A more detailed discussion of the alphabetical arrangement which is used in the book and of the special arrangement of the grammatical categories is given below.

An abbreviation indicates the word class of each lemma. It is followed by two numbers. The former indicates in how many texts (of 100) the lemma occurred while the latter indicates the total number of examples of this lemma.
The following abbreviations are used for the word classes of the lemmas:

- **ao**: adverb
- **erl**: foreign word
- **fn**: pronoun
- **gr**: article
- **lo**: adjective
- **no**: noun
- **ogr**: unanalysed word
- **so**: verb
- **st**: conjunction
- **to**: numeral

**Table I. Word class abbreviations for the lemmas.**

In a few instances the gender of nouns is also given when the word exists in more than one gender. In this case the abbreviation **kk** is used for masculine, **kkf** for feminine and **hk** for neuter. The abbreviations **dfn**, **sfn** and **tfn** are also used to distinguish between indefinite pronouns, interrogative pronouns and relative pronouns where necessary.

As was previously mentioned, formal characteristics are used to distinguish between lemmas. The word class is the most important, as words are classed under different lemmas if the word class is not the same. In addition, the gender of nouns distinguishes between lemmas, as does the type of pronoun. Verbs which have the same reference form but otherwise different inflections are also classed separately (cf. section 4.3).

Word forms are entered under their respective lemmas and each word form is followed by an analytical tag which shows the inflection of the word. The word form together with the analytical tag is referred to as an analysed word form as previously indicated. Each analysed word form is followed by a number which indicates the number of examples of the analysed word form in question in the text corpus of the survey.

In the analytical tag which follows each word form the grammatical aspects are represented by one-letter abbreviations. The first letter of the tag always represents the word class. A noun is represented by **n**, and adjective by **l** etc. The meaning of each abbreviation is dependent upon the "column" in which it stands and a resumé of all the abbreviations which are used in the analytical tags is given in Table II. As was previously mentioned, the first letter of the analytical tag represents the word class. If the first letter is **n** it indicates that the word is a noun. The gender of the noun is indicated in the second column, the number in the third, the case in the fourth and in the fifth column a **g** indicates the word has a suffixed definite article. The sixth letter is used to distinguish between different varieties of proper nouns, (m = names of persons, o = of places and s = other proper nouns).

In attempting to have the abbreviations as transparent as possible we have generally used the first letter of the Icelandic word for the grammatical category in question. This can cause conflicts, however. As the dative (págifull) is represented by the letter **v**, another letter had to be selected for the accusative (póifull) and **o** was chosen.

To return to the extract from the word list which was shown here above, there is one example of the noun **föstra** in the survey. This is the word form **föstra** which has the analytical tag **vn**. With reference to Table II this can be interpreted to mean that the word form is a noun (n), feminine (v) singular (e) and occurs in the nominative (n). There is one example of the verb **föstra** in the word form **föstrar** which has the analytical tag **sgfne3**. This indicates a verb (s) in the active voice (g), indicative mood (f), present tense (n), singular (e) third person (3).
Analysed word forms are not always included with lemmas. Analysed word forms are not given, for instance, for foreign or unanalysed words or for numerals. In these cases only the lemma is entered. In a few special cases explanations in brackets have been added to lemmas when it is not obvious how they should be read; this is the case, for example, when a hyphen is used for *til* (to) (p. 3). Finally it should be mentioned that some words are so long that it has been necessary to divide them between lines. In such a case the hyphen is added to the first part of the word. If there already is a hyphen there, this is moved to the next line along with the latter part of the word and this indicates whether the hyphen is a part of the original word or not. See, for example, the word *tvöpíssund-og-fimmhundruð-krónur* (p. 484).

5.1.2 Order

5.1.2.1 Alphabetical order

If a dictionary user is to be able to find what he is looking for in a dictionary it is absolutely necessary that the words be arranged in a specific manner. They are most commonly arranged in alphabetical order and this practice has been adopted in the chapter on lemmas and inflected forms [1]. In other chapters the words are arranged in order of frequency (e.g. [3] and [7]) or in alphabetical order according to the word endings ([6] and [8]).

At first sight, it might appear to be scarcely necessary to discuss at length the manner in which the words in this book have been alphabetically ordered (although alphabetical ordering has been the subject of learned articles by numerous scholars, cf., for example, Gavare 1988). The manner in which words should be arranged in alphabetical order has become reasonably well established and we have followed the trend, which is now well on the way to becoming generally accepted in Iceland, of ordering the accented Icelandic letters separately (cf. Baldur Jónsson 1987).

The frequency dictionary includes various symbols which are not part of the traditional alphabet and for this reason it was decided to provide a table showing the order which was followed. Generally speaking, the arrangement places non-alphanumeric symbols and numerals first, then alphabetical characters, with accented letters arranged separately following the respective unaccented vowels (e.g. 'á' follows 'a'). Various foreign letters with diacritical marks come between the respective unaccented vowels and the accented ones ('á' thus comes between 'a' and 'â.).

In ordering words there is no distinction made between upper and lower case letters, although if there is no other difference between two words except that one begins with an upper and the other with a lower case letter, the word beginning with a lower case letter precedes that beginning with an upper case one. The word lists are all arranged "in dictionary form", with regard only for alphanumeric characters. Other symbols (such as hyphens, for example) have no effect on the ordering. A word such as *A-bekkur* thus comes between the words *Abdi* and *Aberdeen* instead of between the words *A-t* and *a-d*, where all these words would have followed directly after *A*.

All the sorting was carried out by computer using the *sort* program of the Unix operating system, with switches -f and -d. A few symbols which were not included in the character set of the computer were arranged afterwards manually. These included ö and —.
<table>
<thead>
<tr>
<th>Column</th>
<th>Category</th>
<th>Analytical symbol</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Word class</td>
<td>N-noun</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Gender</td>
<td>K-masculine, V-feminine, H-neuter, X-gender unspecified</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Number</td>
<td>E-singular, F-plural</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Case</td>
<td>N-nominative, O-accusative, P-dative, E-genitive</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Article</td>
<td>G-with suffixed definite article</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Proper nouns</td>
<td>M-name of person, D-place name, S-other proper noun</td>
<td></td>
</tr>
</tbody>
</table>

| 1      | Word class | P-pronoun |             |
| 2      | Subcategory | A-demonstrative pronoun, B-indefinite demonstrative pronoun, E-possessive pronoun, O-indefinite pronoun, P-personal pronoun, S-interrogative pronoun, T-relative pronoun |             |
| 3      | Gender/Pronoun | K-masculine, V-feminine, H-neuter / 1-1st pers., 2-2nd pers. |             |
| 4      | Number     | E-singular, F-plural |             |
| 5      | Case       | N-nominative, O-accusative, P-dative, E-genitive |             |

| 1      | Word class | G-article |             |
| 2      | Gender     | K-masculine, V-feminine, H-neuter |             |
| 3      | Number     | E-singular, F-plural |             |
| 4      | Case       | N-nominative, O-accusative, P-dative, E-genitive |             |

| 1      | Word class | T-numeral |             |
| 2      | Category   | F-cardinal number |             |
| 3      | Gender     | K-masculine, V-feminine, H-neuter |             |
| 4      | Number     | E-singular, F-plural |             |
| 5      | Case       | N-nominative, O-accusative, P-dative, E-genitive |             |

| 1      | Word class | S-verb (except for past participles) |             |
| 2      | Voice      | G-active, M-middle |             |
| 3      | Mood       | N-infinitive, B-imperative, F-indicative, V-subjunctive, S-supine, L-present participle |             |
| 4      | Tense      | N-present, P-past |             |
| 5      | Number     | E-singular, F-plural |             |
| 6      | Person     | 1-1st person, 2-2nd person, 3-3rd person |             |

| 1      | Word class | S-verb (past participle) |             |
| 2      | Voice      | G-active, M-middle |             |
| 3      | Mood       | P-past participle |             |
| 4      | Gender     | K-masculine, V-feminine, H-neuter |             |
| 5      | Number     | E-singular, F-plural |             |
| 6      | Case       | N-nominative, O-accusative |             |

| 1      | Word class | A-adverb |             |
| 2      | Degree     | M-comparative, E-superlative |             |
| 3      | Category/Case governor | A-does not govern case, U-exclamation / O-governs accusative, P-governs dative, E-governs genitive |             |
| 4      | Word class | C-conjunction |             |
| 5      | Category   | N-sign of infinitive, T-relative conjunction |             |
| 6      | Category   | E-foreign word |             |

Table II. Explanations of abbreviations used in analytical tags.
Table III shows the order of all the symbols which occur in the book. The columns are read from top to bottom beginning on the left side.

<table>
<thead>
<tr>
<th>%</th>
<th>8</th>
<th>é</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>%o</td>
<td>9</td>
<td>f</td>
<td>s</td>
</tr>
<tr>
<td>&amp;</td>
<td>a</td>
<td>g</td>
<td>t</td>
</tr>
<tr>
<td>'</td>
<td>á</td>
<td>h</td>
<td>u</td>
</tr>
<tr>
<td>(</td>
<td>á</td>
<td>l</td>
<td>ü</td>
</tr>
<tr>
<td>)</td>
<td>a</td>
<td>i</td>
<td>ü</td>
</tr>
<tr>
<td>+</td>
<td>½</td>
<td>a</td>
<td>j</td>
</tr>
<tr>
<td>+</td>
<td>0</td>
<td>b</td>
<td>k</td>
</tr>
<tr>
<td>±</td>
<td>1</td>
<td>c</td>
<td>l</td>
</tr>
<tr>
<td>,</td>
<td>2</td>
<td>d</td>
<td>m</td>
</tr>
<tr>
<td>-</td>
<td>3</td>
<td>ð</td>
<td>n</td>
</tr>
<tr>
<td>- [to]</td>
<td>4</td>
<td>ð</td>
<td>o</td>
</tr>
<tr>
<td>- [minus]</td>
<td>5</td>
<td>e</td>
<td>ð</td>
</tr>
<tr>
<td>+</td>
<td>6</td>
<td>è</td>
<td>p</td>
</tr>
<tr>
<td>×</td>
<td>7</td>
<td>é</td>
<td>q</td>
</tr>
</tbody>
</table>

Table III: Arrangement of symbols in the frequency dictionary.

5.1.2.2 Arrangement of analytical tags

In the chapter on lemmas and inflected forms the analysed word forms are arranged specifically within each lemma, for the most part following a system of "grammatical ordering" which is based primarily on traditional accidence (cf. Björn Guðfinnsson 1937).

The grammatical categories are arranged in the manner shown in Table II. To begin with they are arranged by the first letter of the tag, then the second, etc. It must be kept in mind, however, that the internal arrangement of the letters does not follow their alphabetical order. To take the nouns for example, lemmas of different genders are kept separate, as was previously discussed. In the next stage of arrangement, the singular precedes the plural (in accordance with the order of the symbols which are used to mark number, ð and ð). With regard to case, the nouns are arranged with the nominative first, followed by the accusative, then the dative and finally the genitive. This arrangement does not follow the alphabetical ordering of the letters which indicate the various cases, as in that case the genitive would come first. By making use of Table II, however, interpreting the arrangement is quite simple.

Attention should be especially drawn to the manner in which the analytical tags for verbs are given. If a verb is in the past participle form the order of categories is as follows: word class, voice, mood, gender, number and case. Otherwise the order is the following: word class, voice, mood, tense, number and person. Voice and mood are, in other words, always given directly following the word class abbreviation for verbs but thereafter the order differs according to whether the word is a past participle or not.
5.2 Most common lemmas in alphabetical order [2] and in order of frequency [3]

5.2.1 Material

In chapters [2] and [3] the frequencies of lemmas are given. In the former the most common lemmas are arranged in alphabetical order, while in the latter they are arranged in declining order of estimated frequency.

For each lemma in [2] the rank, estimated frequency, frequency and dispersion are given, along with the subject areas in which it occurs and in how many texts.

Only some of the aspects given in [2] are specified in [3], i.e. the rank of the lemma, the lemma itself and the estimated frequency.

The statistical concepts used here should be explained briefly.

5.2.2 Distribution and rank of words — statistical concepts


Words and their individual forms are distributed differently in the various subtexts and subject areas. To describe this distribution in a statistical manner and predict what frequency is most likely to occur in any new text which conforms to the selection of texts in the survey, the values $D$, dispersion, and $F_v$, estimated frequency, are used. These values can refer to a word form, lemma, inflected word form or similar forms. The following values are used in calculating the dispersion and estimated frequency:

\[
\begin{align*}
    n & = \text{number of subtexts} \\
    m & = \text{average frequency} \\
    x & = \text{frequency in individual subtexts} \\
    s & = \text{standard deviation} \\
    F & = \text{observed frequency} \\
    D & = \text{dispersion} \\
    F_v & = \text{estimated frequency}
\end{align*}
\]

The calculations are performed in the following manner: If subtexts differ in size, the values of $x$ have to be adjusted to take into account the size of the subtexts, i.e. the number of running words they contain.

The average frequency $m$ is calculated according to the formula:

\[
m = \frac{\sum x}{n}
\]

The standard deviation $s$ is calculated according to the formula:

\[
s = \sqrt{\frac{\sum (x - m)^2}{n}}
\]

Dispersion $D$ is calculated according to the formula:
\[ D = 1 - \frac{s}{m\sqrt{n-1}} \]

Estimated frequency \( F_e \) is calculated according to the formula:

\[ F_e = D \cdot F \]

and this estimated frequency applies for equally large portions of text.

The results of the calculations may be interpreted as follows: If \( D = 1 \) then \( F_e = F \) which indicates a perfectly even distribution. If \( D = 0 \) then \( F_e = 0 \). This is the case if the word occurs in only one subtext. Otherwise \( D \) is always to be found within the range \( 0 < D < 1 \).

By calculating the estimated frequencies of words we are attempting to give a better picture of word frequency than by using only the observed frequency. The frequency of words in a given text has little predictive value for the frequency of words in another text. By using estimated frequency it is, however, possible to estimate the frequency of words independent of the texts which are being examined in each instance.

In all instances the estimated frequency is lower than the observed frequency, which can be explained by the fact that other complete texts include a great number of new words which will be included in the total number of running words.

If the dispersion value is close to 1, and the estimated frequency thus almost equal to the observed frequency, this means that the word is equally common in all the subtexts.

If the dispersion value is, on the other hand, appreciably lower than 1 then the estimated frequency will be much lower than the observed frequency and the word very unequally distributed among the subtexts. If the word occurs in only one subtext both the dispersion value and the estimated frequency will be 0.

In this survey the dispersion and estimated frequency were calculated for all the lemmas on the basis of the five subject areas. The difference in the total number of running words in each of the five subject areas is so small that its effect on the value of the dispersion is negligible. For this reason the values for \( x \) were not recalculated on the basis of the size of text in each subject area as would be required if the previous account were followed to the letter.

The rank of the lemmas is calculated on the basis of the estimated frequency. The word which has the greatest estimated frequency is ranked number 1, the word with the next greatest estimated frequency is number 2, etc. Words occasionally have the same estimated frequency and thus receive the same ranking. The lemmas \( losna \) and \( tölö \), for example, have the same estimated frequency, 37.42. The word preceding them has rank number 728 and these two words thus have the rank 729 and 730. Strictly speaking, their rank should be 729.5, which is the average of 729 and 730. All ranks used in the book are whole numbers, and fractions are lowered to the next whole number and marked with an asterisk. In [2] and [3] the words \( losna \) and \( tölö \) are thus both given the rank 729*.

In chapter [3] the lemmas are arranged in descending order of estimated frequency. It should be kept in mind that the figures for estimated frequency are printed with a varying number of decimal places. This was done primarily for technical reasons in printing. The number of decimal places given for the values of estimated frequency thus does not indicate the exactitude of the values. The degree of uncertainty for these values is, in fact, considerably more than indicated by the number of decimal places.
5.3 Lemmas in individual word classes [4] and subject areas [5]

Chapter [4] includes the most common lemmas of individual word classes. There are a total of eight lists in this chapter, one for each word class. Each list contains the 100 most common lemmas of the word class concerned and, in addition, the 25 most common lemmas of each subject area are listed, even though they are not among the 100 most common lemmas of the survey. The word classes are arranged in the following order: nouns (598–601), adjectives (602–605), pronouns and independent article (606–607), numerals (608–611), verbs (612–615), adverbs (616–619) and conjunctions (620–621).

An examination of these lists reveals a number of points of interest. To take the nouns, for a start, it turns out that the word *madur* (man) is the most common noun if the results of all the 100 texts are taken together. Its rank differs, however, from one subject area to another. In the first and second group the word is ranked in first place, it is in second place in the third group and in third place in the fourth and fifth group. The most common nouns in the third and fourth group are *dr* (year), and *mamma* in the fifth group. The adjective *mikill* (much/great/large) is the most common adjective in the survey as a whole and is also the most common adjective in all the subject areas except the fifth, where it is ranked third. It is perhaps scarcely a coincidence that in the fifth group, books for children and youths, *litil* (little/small) is the most common adjective! There is somewhat more correspondence among the subject areas with regard to the most common words in other word classes. The word *l* ranks first among adverbs in all of them, as does *einn* among the numerals and *og* among the conjunctions.

Chapter [5] lists the 100 most common lemmas in each subject area irrespective of their word class. The verb *vera* (to be) and the conjunction *og* alternate in first rank in the five subject areas.

5.4 Lemmas in alphabetical order according to ending [6]

When words are arranged in dictionaries in alphabetical order it is generally the initial letter which is first taken into account, followed by the second and so on to the final letter. It is then easy to find the words according to their beginnings. For grammarians it can also be useful to arrange the words so that those having common endings are arranged together in word lists. In this manner words which end with a certain concluding sequence, such as *-legur*, *-heit* etc. are grouped together. This has been done in chapter [6], Lemmas in alphabetical order according to ending.

In chapter [6] it is easy to look up lemmas according to their endings. For each word the observed frequency is given for all the subject areas combined. When using this list the main difference is that in order to follow the alphabetical arrangement the words must be read “from right to left”. For this reason the words are set flush-right in columns as is often done when numbers are arranged in columns.

| fjævera | no | 5 |
| vistarvera | no | 9 |
| nævera | no | 5 |
| sēra | no | 45 |
| marsēra | so | 3 |
| grassēra | so | 1 |
| spåssēra | so | 1 |

This arrangement is intended to make it easier for users to find the word endings in which they are interested. As the example here above shows, the word class and frequency of each word is also specified.
5.5 Word forms [7] and [8]

The next two chapters deal with word forms. There were a total of 59,343 word forms in this survey. In the former of the two chapters, The most common word forms in order of frequency [7], word forms are arranged in order of estimated frequency. The most common word form turned out to be og.

The latter chapter contains information on the word forms in alphabetical order by ending [8]. This list is very similar to the corresponding lists of lemmas [6] except for the fact that there is no word class specified as word forms belonging to different word classes are grouped together. In other respects the list is self-explanatory.

5.6 Grammatical ambiguity of word forms [9]

The mechanical analysis used with the grammatical classification of the running words was discussed briefly earlier. Mention was made of the fact that the success of the analysis was approximately 80%, which must be considered as relatively good considering the fact that this is the first version of the program. But why is it so difficult to achieve an analysis which is 100% correct? The primary reason is the grammatical ambiguity of the running words, as it is not even always possible to determine from the word alone whether it is a noun or verb, not to speak of determining definitively its case, person or number (to mention some examples). This factor was investigated especially and in chapter [9] information on this “grammatical ambiguity” of word forms appears.

In order to collect information on this question, each individual word form was investigated in order to discover how many different analytical tags could be attached to it. The word form minni exhibited the greatest variety, having 24 different analytical tags in this survey. Of these, seventeen were adjective tags, five were noun tags, one a pronoun tag and one a verb tag. This means that if an attempt is made to classify the word mechanically when it occurs in a text, the program has at least 24 different possibilities to choose from. To achieve reasonable success, mechanical analysis must make reference to the context in which words occur. The following table indicates how many tags were attached to individual word forms. The table shows that, of the 59,343 different word forms in the survey, 9,441 or 15.9% had more than one analytical tag and are thus grammatically ambiguous (in addition to which a fair number of those which had only one analytical tag would doubtless have also proved to be grammatically ambiguous if there had been more examples of word forms in the survey). All the word forms which had more than four analytical tags are listed in chapter [9].

<table>
<thead>
<tr>
<th>Number of word forms</th>
<th>Number of analysed forms</th>
<th>Number of word forms</th>
<th>Number of analysed forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>26</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>69</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>96</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>209</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>579</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>1,772</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
<td>6,586</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>49,902</td>
<td>1</td>
</tr>
</tbody>
</table>

Table IV. Summary of grammatical ambiguity.
5.7 Graphemes, digrams and trigrams [10], [11] and [12]

The next three chapters deal with the frequency of graphemes, digrams and trigrams. In counting the graphemes no distinction is made between upper and lower case letters and upper case letters are thus included in the frequency of the letters.

Spaces between words and sentence-level punctuation are not counted, only those symbols which are parts of, or connected to, words.

In this manner a total of 75 different graphemes was obtained. Three sorts of frequencies are given, for running words, word forms and lemmas. The results are shown side by side in lists to facilitate comparison.

Comparison reveals, for instance, that words containing the letter þ occur relatively often, as the proportional frequency of this letter in running words, 1.57%, is considerably higher than in word forms and lemmas, 0.46%.

In other tables the graphemes are arranged in order of frequency and, in addition, the frequency with which graphemes occur at the beginning and end of words is given. From this it can be seen, and should scarcely come as a surprise, that the letter ð never occurs in word-initial position and the letter þ only occurs seven times in word-final position. The most common symbols at the beginning of lemmas are s (14.54%) and l (9.33%) which concurs with the amount of space occupied by words beginning with these letters in dictionaries.

Digrams and trigrams are two- and three-letter combinations, the frequency of which was investigated separately in running words, word forms and lemmas, and also with respect to their position in words. In order to determine digrams in a given word each two consecutive graphemes are combined, beginning from the left edge of the word and proceeding to the right. The following digrams, for example, are contained in the word tvistofungar:

'tv', 'vi', 'is', 'st', 'tö', 'öf', 'fu', 'un', 'ng', 'ga', 'ar'

and the following trigrams were determined in the same fashion:

'tvi', 'vis', 'ist', 'stö', 'töf', 'öfu', 'fun', 'ung', 'nga', 'gar'

It should be mentioned that only those combinations which consisted of two or three letters were considered. In counting the di- and trigrams all the upper case letters were converted to lower case. If the number of letters in the Icelandic alphabet is considered as 36 then the total number of digrams possible is 36 × 36, or 1,296. Of these, 916 were present in the survey material, or 70.7%. The total number of trigrams possible is 36 × 36 × 36 or 46,656. Of these, 10,141 or 21.7% occurred. The frequencies of digrams is given in chapter [11] where all the digrams are listed. The list also includes several digrams with one or more letters which do not occur in the Icelandic alphabet, letters such as 'ü' and 'ö'. A total of 30 such di- and trigrams occurred in the survey. The frequencies of trigrams are given in chapter [12]. Because of the large number of trigrams, however, only those with a relative frequency greater than 0.02% are listed.

In closing, mention may be made of an informative article on the use of trigrams in searching for spelling errors in computerized text in English (MacMahon, Cherry and Morris 1978).
5.8 Analytical tags [13]

The next chapter provides an overview of all the analytical tags which occur in the survey. They proved to be 621 in number. Their frequency, however, varies greatly. The most common analytical tag proved to be A-A, which was attached to 50,282 running words. This was followed by c, of which there were 42,491 examples; nineteen analytical tags occurred only once each. The analytical tags are given both in alphabetical order and in order of frequency.

5.9 Summary tables [14]

The final chapter of the book contains a number of tables containing information regarding the survey in its entirety and individual word classes. The material of the tables is explained in individual explanatory notes accompanying each and it is unnecessary to discuss them further here.
7 Bibliography

7.1 Works referred to in the introduction


7.2 Other works consulted


