An Estonian morphological analyser and the impact of a corpus on its development

Heiki-Jaan Kaalep
University of Tartu
Tiigi 78 Tartu Estonia EE2400
hkaalep@psych.ut.ee

Keywords: computer implementation, Estonian, language engineering, morphology, text corpora

Summary

The paper describes a morphological analyser for Estonian and how using a text corpus influenced the process of creating it and the resulting program itself. The influence is not limited with the lexicon only, but is noticeable in the resulting algorithm and implementation too. When work on the analyser started, there was no computational treatment of Estonian derivatives and compounds. After some cycles of development and testing on the corpus, we came up with an acceptable algorithm for their treatment. Both the morphological analyser and the speller based on it have been successfully marketed.

1. Introduction

The increased use of personal computers throughout the world is bringing with it a demand for language technology products, starting with spell-checkers, for languages inadequately described by computational linguistics. With few resources to spend on basic scientific research today, are these languages doomed to the backyard of language technology? Fortunately, text corpora can in some cases compensate for the lack of highly qualified computational linguists and basic linguistic research. This optimism derives from our experience of creating a morphological analyser and speller for Estonian in an iterative process of validating and remaking the algorithm, based on results from processing a corpus.

The situation in Estonian computational linguistics in 1991, when work on creating a morphological analyser and speller started, was the following. A machine-readable dictionary of simplex words, superbly suitable for treating inflection, was available. But very little was known about the usage of Estonian in real texts, and there had been no attempts to treat Estonian derivation and compounding computationally, even though derived and compound words comprise up to 20% of the word tokens in Estonian texts. So, an algorithm for these phenomena had to be devised from scratch, checked for validity and effectiveness, and, in some cases, completely rewritten.

Creating a morphological analyser for Estonian was motivated by the need for a speller. Because the goal was a commercial product, issues of implementation and effectiveness were important from the very beginning.

Using a text corpus provided a basis not only for updating the lexicon (treated only superficially in the current paper), but also for the design and implementation of the algorithm of analysis, an issue that has been virtually neglected in corpus linguistics.

In this paper, only the morphological analyser ESTMORF is described, and not the speller, because the two were developed simultaneously and they are algorithmically identical. The speller is different only in that it does not lemmatise and does not find multiple morphological readings of a word form.

2. The Corpus

The corpus we used during the development of the analyser and speller consisted of several different corpora: the Corpus of the Estonian Literary Language (CELL), the Corpus of Baltic News Service On-line News (CBNS) and the Corpus of Estonian Newspapers (CEN).

2.1. CELL

The project to develop the Corpus of the Estonian Literary Language (CELL) was launched in the Laboratory of the Estonian Language, University of Tartu, in autumn 1991 [Hennoste et al. 1993]. CELL was designed following the principles of closed corpora such as the Brown corpus [Francis et al. 1964], the Lancaster-Oslo/Bergen or LOB corpus [Johansson et al. 1978], and the London-Lund corpus [Svartvik et al. 1980]. The planned 1 million word volume of the corpus was achieved in spring 1995.

When choosing the texts for CELL, it was considered important that the texts should represent the cultural situation of the chosen time frame (years 1983-1987). The 1 million words in CELL are composed of about 500 texts, each containing 2,000 words. LOB was based on the same division and contains texts from only one year. The number of Estonian literary texts is so small that if texts produced in a single year were used, the extracts would have had to be much longer than in LOB. Therefore a five-year period (1983-1987) was chosen, during which language usage was rather stable. Also similar to LOB, the corpus texts are divided into different text classes (fiction, press, etc.). A different number of texts has been chosen from every class. The decisions were based upon the statistics of the English and American corpora and on the opinions of experts about the number, dissemination and influence of texts in various areas of Estonian culture. The following table summarises the text classes and their percentages from the overall volume in LOB and CELL.

No.	Text class	LOB	CELL
1.	Newspapers	17,5	17,5
2.	Religion	3,5	0,8
3.	Hobbies	7,5	7,5
4.	Popular lore	9,0	15,5
5.	Biographies, essays	15,0	9,0
6.	Documents	6,0	1,2
7.	Science	16,0	15,5
8.	Fiction	25,0	25,0
9.	Encyclopaedia	-	2,0
10.	Propaganda	-	6,0
		99,5 %	100 %

Table I. Text classes and their percentages from the overall volume in LOB and CELL

From the very beginning it was decided that CELL should be tagged in order to be of more use for linguists. The tagging of CELL follows the guidelines of TEI [Guidelines 1994]. All the newspaper texts, 175 thousand words, have been tagged for paragraphs, sentences, numbers, abbreviations and acronyms, proper names, direct speech, quotations and non-literary Estonian. The rest of the corpus has been so far tagged for typographical changes (e.g. boldface and italics), paragraphs and sentences. The tagging and validation were performed manually with SGMLS.

Actually, we have found that tagging was of no help for developing a morphological analyser or speller. Mark-up may have an impact on the development of a morphological analyser in two ways:

1. It may simplify the analysis, allowing skipping proper names, abbreviations, acronyms and other tokens normally not found in dictionaries and thus not included in the lexicon of the analyser. However, this would be an over-simplification for a program designed to analyse unrestricted text.
2. Mark-up may simplify the task of devising the algorithm, by providing correct solutions in the tagged texts. In case of Estonian, however, the manually pre-tagged texts provided little additional information. It was more rewarding to run the morphological analyser on a plain text and then automatically filter out the proper names (they have an initial capital letter), acronyms and numbers from the unknown tokens. The remaining unknown tokens then served as an input for lexicon building and algorithm devising.

We used a version of the corpus where the mark-up had been deleted. This gave a more realistic environment for the operation.

During the initial phases of developing ESTMORF, we used primarily fiction and newspapers, comprising altogether about 300,000 words then. First, these were the first text classes typed in during the building of CELL, and thus they were available early on. Second, literature is closest to "natural Estonian" and thus the best source for inspiration and testing ESTMORF, while newspapers represent language produced by people who work under strict time constraints and are thus most interested in using a speller. (Remember that ESTMORF and the speller were developed hand in hand and greatly motivated by potential market needs.)

During later phases of developing ESTMORF we used the entire 1 million word corpus for testing.

CELL had the greatest impact on the contents of the lexicon (simplex and compound words and proper names) and on the design of the algorithm.

2.2. CBNS

The Corpus of Baltic News Service On-line News (CBNS) was launched in October 1994. It contains news produced by the news agency BNS and sent to subscribers via e-mail. The texts we receive are archived automatically. The corpus grows steadily with the speed of 3-4 million words a year. The texts are tagged in no way except for the beginning and end of each news-text. Texts produced in each month are archived in one file.

This corpus contains many spelling errors, so it was used for checking the speller on unedited texts. A set of 1000 incorrectly spelt words was extracted, in part manually, from the production of one month of BNS. This set served as a valuable source for determining the typical errors done by native speakers of Estonian, which in turn was very useful in designing the algorithm for suggesting correct forms in replace of misspelled ones.

CBNS was also very useful as a source for proper names, abbreviations and acronyms. But because the collection of texts started relatively late, CBNS was not used in the initial phases of program development and thus influenced only the lexicon, not the algorithm for morphological analysis.

Test runs on 500,000 words were used during the development time.

2.3. CEN

The Corpus of Estonian Newspapers (CEN) was launched in 1993, with the aim of following the changes in language usage over time. It contains texts from various newspapers from the period of perestroika (1989 and 1991) and independence (1993, 1995 and 1996). Currently, CEN contains 4 million words, and since the beginning of 1996 has grown at the rate of 4 million words a year. The mark-up of the texts varies across periods and newspapers.

We used a version of the corpus where the mark-up had been deleted. Only texts from 1989 and 1991 in CEN, altogether 100,000 words, were used in the development of ESTMORF.

3. Estonian morphology

Estonian is usually considered to be an agglutinative language, thus belonging to the same group as Finnish, Turkish, Quechua or Swahili. Estonian contains words of considerable complexity, and parsing such word structures for correctness and structural analysis necessitates a thorough morphological analysis. Words contain no direct indication of where the morpheme boundaries are.

3.1. Inflection

Estonian words can be divided into three main inflectional groups:

1. Declinable words that can change in case and number, e.g. nouns, adjectives etc.
2. Conjugable words, i.e. verbs that can change in mood, tense, voice, person, number, negation, infiniteness and case.
3. Uninflected words.

These three groups can be divided into smaller units, depending on syntactic and/or semantic properties. There is no one and correct classification scheme in this respect; e.g. [Valgma 1970], [Viks 1992] and [EKG 1995] all give different classifications.

ESTMORF differentiates among the following word classes or parts-of-speech:

1. Declinable words: common nouns or substantives (S), proper nouns (H), adjectives with a positive degree (A), adjectives with a comparative degree (C), adjectives with a superlative degree (U), cardinal numerals (N), ordinal numerals (O), pronouns (P) and abbreviations and acronyms (Y). (It is possible in Estonian to glue an inflectional affix to a number, an acronym or abbreviation; e.g. 1995ndal ‘in 1995’, USAs ‘in the USA’, lk-lt ‘from page’.) These make up 27 000 words in the lexicon.
2. Conjugable words: verbs (V); 7000 words.
3. Uninflected words: some uninflected adjectives (A), genitive attributes (G), adverbs (D), adpositions (K), conjunctions (J), interjections (I), some words met only together with certain verbs (X) and the so-called sentence marks (Z). These make up 5500 words in the lexicon.

Word classes Y and Z were added to ESTMORF after testing it on the corpus. A morphological analyser has to attach some kind of interpretation to every word-like unit in a text; and units that should be classified as abbreviations, acronyms or non-words make up over 2% of the word-forms of a running text.

Slight differences exist in the description of the Estonian inflectional system in different sources (cf. [Valgma 1970], [Viks 1992] and [EKG 1995]). ESTMORF follows [Viks 1992], according to which an Estonian declinable word paradigm contains 29 slots (15 cases in singular and 14 in plural), and an

Estonian verb paradigm contains 83 slots. There is not a one-to-one correspondence between paradigm slots and inflectional forms: some slots have different parallel inflectional forms while some inflectional forms are homonymous inside a paradigm. This discrepancy is quite frequent, e.g. thousands of declinable words have paradigms with two parallel plural forms for 12 cases, thus adding 12 word forms to their paradigms, while thousands have homonymous forms for singular nominative, genitive, partitive or additive case. An Estonian verb paradigm has (with a few exceptions) only 47 different word forms to cover 83 slots in a paradigm, leaving 23 word forms ambiguous in two or more ways. ESTMORF uses underspecification to control ambiguity in the verb paradigm. Following [Viks 1992], ESTMORF uses special codes to mark the 47 different word forms of a verb paradigm, e.g. *d* for indicative present active affirmative 2nd person singular, *ksid* for conditional present active affirmative 2nd person singular and 3rd person plural.

Estonian inflection involves appending affixes to a stem, as well as alternations in the stem itself. Every simplex word form consists of two parts: the word stem and the inflectional formative (including zero-morpheme), both of which can vary. Following [Viks 1992] ESTMORF treats inflexional affixes as unstructured ones, without distinguishing single morphemes in their composition, e.g. maja[le 'to a house', maja[de 'houses'', maja[dele 'to houses'. As a rule an Estonian word has more than one stem variant, e.g. padi[0 'pillow', padja[s 'in a pillow', patja[des 'in pillows'.

3.2. Derivation

Derivation, a frequent and productive way in Estonian for forming new words, is a process where adding an affix produces a new morphological word having its own inflectional paradigm. Whether the lexical meaning of the word used as the derivational base remains unchanged is not important for the morphological analyser. In Estonian, derivation is mainly a process of appending derivational suffixes, more than 60 altogether, to both declinable and conjugable words. Suffixes can be appended sequentially; up to four suffixes in a row can be appended in some cases.

Prefixes play a smaller role in Estonian derivation. [EKG 1995] lists 16 prefixes, 14 of which are met mostly in loan words. Prefixes are, as a rule, not used sequentially.

About 8% of the word forms in a running Estonian text are derived words; in journalism and scientific texts the figure is even higher.

3.3. Compounding

In Estonian, compounding is even more frequently used for word formation than derivation. Compound words comprise to more than 12% of running words in an average Estonian text, and even more in newspaper texts.

The formation of Estonian compounds is quite free: inflected words, stems, truncated stems or derived words belonging to any word class (excluding conjunctions and acronyms) may be glued together to form new compound words, although not all combinations are allowed. As a rule, the finite forms of verbs are not compounds, though there are exceptions, e.g. abielluma 'to get married'. There is also a limitation to the number of component stems: there are no examples of words with more than 5 stems [Kask 1967:46]; and rather few with 5 stems, e.g. all+maa+raud+tee+jaam 'subway railway station', raud+tee+üle+sõidu+koht 'railway crossing'. Too long or clumsy-looking compounds are preferably written with a hyphen, e.g. avalik-õiguslik 'public and legal'.

4. ESTMORF, a morphological analyser for Estonian

ESTMORF is a computer program for analysing unrestricted Estonian text. It can be accessed via the Internet (http://www.filosoft.ee/ and follow the links). ESTMORF is implemented in a most straightforward way: it compares word forms of the running text with combinations of lexemes from its lexicon. The comparison involves only the literal comparison of strings. No two-level rules [Koskenniemi 1983] are used.

The main properties of ESTMORF are the following:

1. ESTMORF accounts for written, not spoken, Estonian.
2. Inflectional morphology is treated completely, down to the very last exception.
3. The lexicon contains the stems of the simplex words belonging to the core vocabulary of Estonian, as well as the most frequently used proper names, abbreviations and acronyms. Productively formed compounds and derivatives are not, as a rule, part of the lexicon.
4. Derivations and compounds are analysed algorithmically, thus eliminating the need to list most of Estonian compounds and derivations in the lexicon and enabling proper analysis of new compound and derived words.
5. The algorithm of analysing derivations and compounds is devised so that it finds the most likely combination of component parts for any given word.
6. The analysis is based on dictionary look-up and involves no heuristics. If a word cannot be analysed deterministically, ESTMORF makes no educated guess.
7. ESTMORF itself takes care of the treatment of punctuation and compound proper names.
8. Besides the treatment of the compounding and derivational processes, ESTMORF has no claims to originality in treating Estonian morphological system.
9. Adequate morphological descriptions are assigned to about 97% of tokens in a running text. The remaining 3% that are not analysed are rare words such as proper names, abbreviations, acronyms, specific terminology, slang etc.
10. ESTMORF is a tool for morphological analysis, as well as various more or less practical purposes.
11. ESTMORF does not take into account syntactic and semantic properties such as valency, animatedness or transitivity.

4.1. Output of ESTMORF

ESTMORF determines, for every input word form, the structure of the word (e.g. stem, derivational suffix, inflectional affix), the word class and inflectional categories (e.g. number and case). The abbreviations for inflectional categories are explained at the Internet site for using ESTMORF (http://www.filosoft.ee/) and are compatible with [Viks 1992]. In the examples below, the input word is at the far left margin. The output analyses by ESTMORF, one or more, are indented on separate lines below the input word.

If a word form is ambiguous inside its paradigm, then the possible sets of inflectional categories are given on the same line. The following analysis of kasvataja ‘governess’ shows that the form is ambiguous for aditive, singular genitive and singular nominative. The brackets around (adt) mean that “aditive” is a doubtful reading by [Viks 1992].

kasvataja
    kasvataja+0 //_S_ (adt), sg g, sg n, //

If a word has more than one allowable structure, lemma or word class, then the analyses are displayed on separate lines:

lood
    lood+0 //_S_ sg n, // (plummet)
    lood+d //_S_ pl n, // (limestone regions covered with thin soil and stunted vegetation)
    loog+d //_S_ pl n, // (mown grasses)
    lugu+d //_S_ pl n, // (tales, stories)
    loo+d //_V_ d, // ((you) are creating)

If the word is a derived one or a compound, then:
1. The stem is separated from the previous component by "_".
2. The inflectional affix is separated from the previous component by "+".
3. The derivational suffix is separated from the previous component by "=".

Only the rightmost component is lemmatised.

Examples:

alleshoidmine
    alles_hoid=mine+0 //_S_ sg n, // (preserving)
3aastast
    3_aastane+t //_A_ sg p, // (3-year old)
lastekodukasvataja
    laste_kodu_kasvataja+0 //_S_ (adt), sg g, sg n, // (foundling hospital governess)
elamisväärseid
    ela=mis_väärne+id //_A_ pl p, // ("living-worth", i.e. worth living)
vettehüpe
    ve+tte_hüpe+0 //_S_ sg n, // ("into-water-jump", i.e. plunge, dive)

In foreign proper names consisting of more than one word, like New York, only the last word inflects, e.g. singular inessiv <u>New Yorgis</u> 'in New York'. Such names are treated in an ad hoc manner as compounds; the blank is retained as a separator:

New Yorgis
    New York+s //_H_ sg in, //

4.2. Morphological ambiguity

A word form may be ambiguous for two reasons:

1. There are multiple ways of dividing the word into lexemes, as in:

kapsas
  kapsas+0 //_S_ sg n, // (cabbage, singular nominative)
  kapsas+s //_S_ sg in, // (in cabbage, singular inessive)
  kapsa+s //_V_ s, // (skipped, indicative active imperfect singular 3rd person)

2. The lexemes are similar but may be interpreted in several ways, as in:

lisasid
  lisa+sid //_S_ pl p, // (appendixes, plural partitive)
  lisa+sid //_V_ sid, // (added, indicative active imperfect singular 2nd or plural 3rd person)

Besides several ways of dividing a word into a stem and an inflectional affix (as with kapsas above) or more generally, into several lexemes for derived and compounded words, there is often also more than one way for determining the lemma of the word, as in the case of soe:

soe
  soe+0 //_A_ sg n, // (warm, singular nominative)
  susi+0 //_S_ sg g, // (wolf, singular genitive)
  suge+0 //_V_ o, // (to comb, imperative)

Sometimes an inflectional affix of a word can be interpreted in several ways inside the word paradigm, as sid with verb lisasid above. If the affix is ambiguous in the paradigms of all the words belonging to the same word class, like sid with verbs, then one might underspecify the morphological categories related to that affix. This is what ESTMORF does with several inflectional affixes of verbs, following [Viks 1992]. However, one should not use underspecification if an inflectional affix is ambiguous in the paradigms of only some words, as d, normally used for plural nominative only, is for

ideed
  idee+d //_S_ pl n, sg p, // (idea, plural nominative or singular partitive)

The output of ESTMORF shows that 45% of words in CELL are morphologically ambiguous. This is a high figure for an inflectional language. It is even more notable in view of the ESTMORF notation for verb inflection that substitutes inner-paradigm ambiguity for underspecification and the ESTMORF algorithm for analysing derived and compound words that does not output unlikely lexeme patterns if more likely ones are applicable.

4.3. Implementation

The near standard for computational morphology at present involves using two-level rules and left-to-right or root-driven analysis of input words (e.g. [Sproat 1992]). These devices have been implemented in various morphological analysers for typologically different languages, too numerous to be listed here. ESTMORF, however, uses no two-level rules and words are analysed right-to-left, using affix stripping. In doing so it belongs to a class of analysers, used for several agglutinative languages before.

Several morphological analysers of Russian [Itogi 1983] proceed by stripping affixes off the word, and then attempting to look up the remainder in the lexicon. Only if there is an entry in the lexicon matching the remainder and compatible with the stripped-off affixes is the parse deemed a success. Brodda and Karlsson used affix stripping for analysing Finnish, but without any lexicon of roots [Brodda and Karlsson 1980]. Suffixes were stripped off from the end of the word until no more could be removed, and what was left was considered a root. Proszeky and Tihanyi describe a method, similar to those described in [Itogi 1983], for analysing Hungarian [Proszeky and Tihanyi 1992].

The reason why ESTMORF follows the older path in implementing a morphological analyser lies in the original motivation: to come up with a “black box” as a tool for morphological analysis and spelling. From this language engineering point of view, the exact mechanism for treating inflection was really no issue, as long as it provided correct analyses. It was extremely handy to convert the “Concise Morphological Dictionary of Estonian” (CMD) by Ülle Viks [Viks 1992] that was used as the basis for the lexicon into a form that could be readily used in affix-stripping morphological analysis, without a need to formulate two-level rules. The issues that got more attention in creating ESTMORF were:

1. How adequate is the lexicon for handling Estonian vocabulary in real texts?
2. How should one handle derivations and compounds?
3. What tokens exist in Estonian texts besides ordinary words, and how should one handle them?

Answers to these questions determine the usefulness of the analyser and speller in the end, and answering them took most of the time in developing ESTMORF.

The speed of the program was also considered important from the very beginning because a precise computational tool that does not run in reasonable time is unacceptable. Because there is no indication as to where the morpheme boundaries are, the analyser must make guesses and check in various lists until an acceptable analysis of the word structure is achieved. One could imagine that the right-to-left analysis should be more efficient as the lists of affixes, being smaller than stem lexicons, permit quicker retrieval. It is sensible to look for the stem in the lexicon only after the suitable affixes are found. But on the other hand, there is plenty of evidence that left-to-right parsers are very efficient ([Karlsson 1992], [Solak and Oflazer 1993]). In addition to excellent engineering (fast searching in the lexicons, good data compression, choosing the right programming language and hardware) one might also consider some statistical properties of real-life texts that should be taken into account, and in particular the proportion of words with certain structure in a text. This is where corpora prove to be useful in designing efficient algorithms, and in the process of creating ESTMORF we have sometimes rewritten parts of the program after analysing the statistical properties of words in real texts.

ESTMORF consists of two parts: the program and a set of various lists, the largest of which is the lexicon of stems. The program itself takes up approximately 115 Kbytes, the lists together, about 650 Kbytes.

ESTMORF has been implemented in C under DOS and UNIX. It runs on a PC XT as well as on a Sun Workstation; on the latter with a speed of 700 words per second. Compiling the subroutines of ESTMORF into the Estonian speller for Microsoft Office 95 showed that the program ran as fast as the English speller although Estonian is morphologically more complex than English.

ESTMORF is also used as the lemmatiser in a text-searching module in the database of the Regulations of the Estonian Government, implemented in a textual database TRIP in the State Chancellery of Estonia.

4.4. Stages in creating ESTMORF

The stages of developing and testing ESTMORF are outlined below. Mistakes and bugs in the algorithm, lexicon and program modules were found and corrected in all the stages of development. By reporting deficiencies, independent users of ESTMORF and a speller based on it also helped in developing ESTMORF.

Work on creating ESTMORF started in August 1991 when we obtained a machine-readable version of CMD by Ülle Viks [Viks 1992]. CMD contains about 36 000 Estonian simplex words with full descriptions for generating all the word forms of paradigm. An analyser for simplex words was created in 4 months, from August 1991 to December 1991.

The analyser was able to analyse 75% of words in a running text. This was the starting point for creating an algorithm for analysing derivations and compounds. We did not know answers for the following questions:

1. How productive are deriving and compounding in real texts?
2. What are the patterns for derivation and compounding and which of these patterns are productive?

Previous work by [Kask 1967], [Kull 1967], [Kasik 1984] and [Kasik 1992] contained useful hints for answering both questions, but could not be implemented immediately. Additional difficulties arose from the tradition of describing derivation and compounding as a process of synthesis, while we were interested in analysis. A separate problem was that compounding had been described as a process of concatenating two components while real texts contain compounds with up to five components. It was unclear to what extent one may apply for more complex compounds recursively the same rules as are applicable for two-component compounds.

In order to be able to create an algorithm for analysing compounds we divided the task into two sub-tasks:

1. Find what are the structural patterns of compounds in real texts.
2. Find the constraints limiting the use of any single pattern.

We examined every structural pattern separately and tried to be very restrictive with constraints, so as to allow only the definitely correct combinations. As an example, consider patterns stem1+stem2 and stem1+stem2+stem3. At the beginning we presumed that stem1 and stem2 are more free to combine into a two-component structure than into a three-component one, given that there are fewer instances of three-component compounds than two-component ones. When we later found by testing that we had been too restrictive, we relaxed the constraints. Every time we changed the set of analysed structural patterns and constraints, we tested the result on the same texts as before. The amount of unrecognisable words diminished step by step. When it was small enough, we tested on new texts, and the cycle repeated.

We finished the development of our algorithm for parsing the derived and compound words when we came to the point where:

1. the amount of unrecognised derived words and compounds in a new text was about the same as the amount of simplex words, and
2. the simplex words were so exceptional that they should not be included in the lexicon.

We concluded that the situation was similar to simplex words, and that apparently these unrecognised compounds represented infrequent methods of word formation in Estonian that could be classified as “exceptional” or “non-orthological”, and which therefore should not be allowed in the algorithm.

In addition to maximising the set of recognised words, we kept an eye on the speed of the algorithm. In order to minimise the time that is wasted on trying to impose false structural patterns on strings, we organised the program so that the most likely patterns were tried first.

If a compound word represented a rare pattern, it was easier to include it in the lexicon than to modify the algorithm in a special way.

The first stage of developing algorithm for derivations and compounds started in January 1992 and continued until 1994. As a result, a speller was created and released to independent users. The corpus used was small: 100,000 words of literature texts and various small in-house texts (articles, letters etc.). The structure of every unanalysed word in the corpus was carefully evaluated independently of the statistical data from the corpus to determine how typical and natural it was, and the algorithm was changed only where such typical patterns were judged to exist.

Initially, ESTMORF presumed that only two-component compounds are freely allowed. To be able to analyse more complex compounds, two lists were introduced: stems that can be concatenated to the beginning and to the end of the word. Thousands of irregular compound words were also added to the lexicon, the reason for irregularity being the non-existence of a component as a separate word or the component belonging to a word class normally not participating in compounds. The initial source for the

lists and irregular compounds was [Viks 1992]. The algorithm for analysing compounds did not make a clear distinction in the analysis of different structural patterns. For example, it would be sensible to first check all the possible ways for a word to have the structure stem1+stem2, and only in the case of failure check for stem1+stem2+stem3. ESTMORF, however, first tried to find some component at the beginning of the word and then parse the remaining part of the word at any cost, in order to find some allowed combination of stems and affixes; and only in the case of failure was a new first component tried.

In 1994 we considered ESTMORF mature enough to be used as a tool in quantitative linguistic studies. We analysed a 300,000-word sub-corpus of fiction and newspaper texts from CELL, and a 100,000-word sub-corpus from 1989 and 1991 in CEN. We found that ESTMORF did not recognise 4% of words in fiction and 9% in newspaper texts, predominantly proper names, abbreviations, acronyms and strings containing numbers. So, thousands of proper names had to be added to the lexicon and the algorithm for treating numbers and other non-words added to ESTMORF.

It also appeared that the algorithm was not optimal for the structural patterns of words in real texts. At first, ESTMORF started from the longest affixes, assuming that if an affix is allowable, then it is very likely that the remainder is an allowable stem, and thus it minimises the need for dictionary look-up. During tests we found, however, that more than half of all the word forms in a running Estonian text are either words without inflectional affix, e.g. et 'that' or have a 0-affix, e.g. raha 'money'. Typically, simplex words make up 75-85% of all the words in an Estonian text. So if we start the analysis from just looking up the word form from the lexicon, without any affix stripping, we will get a positive result for 40% of all the word forms with the first try. Based on the data we changed the algorithm for analysing simplex words. ESTMORF now starts by stripping the shortest possible affix. Actual test-runs with alternative ways of affix stripping showed that starting from the shorter ones resulted in faster spell checking.

In compound word analysis we clearly separated the modules for treating words with different structures and reordered them, so that more likely structural patterns are tried first.

In 1995 we tested ESTMORF on one month's news texts from CBNS, 500,000 words altogether. As a result more proper names were added to the lexicon. It also appeared that the texts contained many spelling errors. A list of 1,000 incorrect words was used for improving ESTMORF. Imitating a suggesting module of a speller, we generated new words from the incorrect ones and checked their correctness. ESTMORF was not restrictive enough in rejecting weird words, so we decided to add one more list to be used by the analyser: a list of stems, not allowed as a component of a compound word. Every time ESTMORF finds a possible component while analysing a compound word it checks if the component is not in this "black list" of stems.

In 1996 we lemmatised a 300,000-word corpus of the legislative directives of the Estonian government from 1995. As a result, more names, mostly of Russian origin, were added to the lexicon. For another large-scale experiment we analysed the whole 1-million word CELL. It appeared that 15,000 simplex words from the ESTMORF lexicon of 32,000 simplex words were never encountered in the corpus. We checked the 15,000-word list by hand and deleted 1800 obsolete and dialect words from the lexicon.

In 1996 an analysis of G. Orwell's "1984" (79,000 words) finally showed that ESTMORF can be considered more or less complete. Only 2% of words were left unanalysed, including mostly British proper names and Newspeak words.

5. Analysing simplex words

The analysis of a simplex word is a cycle of inflectional affix stripping, dictionary look-up and partitioning correctness checking:

First, cut an inflectional affix from the end of the word, then check if the first part of the word can be found in the lexicon of word stems and then check if the stem and the affix fit together. E.g. ütelda 'to say' consists of the stem ütel and an inflectional affix da. The checking of compatibility is necessary in order to filter out words like ütelta which consists of a normal stem and an affix, incompatible in this particular case.

The affix stripping in ESTMORF starts from the shortest possible ones. The longer an inflectional affix is, the less instances of it we find in texts and so the less probable it is that any given word has it.

6. Analysing derived words

ESTMORF uses a list of 40 productive derivational suffixes which can be appended to substantives, adjectives, numerals or verbs, resulting in substantives, adjectives or adverbs. Some suffixes can be appended to only one word class, some to several different ones, resulting in different word classes. The constraints for derivation involve the word class of the stem, the form of the stem (e.g. sometimes a stem of singular genitive may attach a derivational suffix, but stem of a singular nominative cannot), and the ending letters of the stem.

For example, -dus can be appended to a verb in participle past passive affirmative, resulting in a substantive as in töödeldud: töödeldus 'processed: processedness', or -dus can be appended to a stem of a singular genitive of an adjective ending with -eda by substituting -eda with -edus as in müreda: müredus 'stale (milk): staleness (of milk)'.

Many derivational suffixes can combine. E.g. -ja and -lik give -jalik, as in pusklema 'to be butting each other', puskleja '(s)he who is butting the other', pusklejalik 'like someone who is butting the other'. ESTMORF does not allow the derivational suffixes to recurse; it uses a list of more than 100 combined suffixes instead. The constraints for appending combined suffixes and the resulting word classes are similar to those of simplex derivational suffixes.

Traditionally, Estonian has been described as having very few prefixes: only eba- (negation) and mitte- (negation) for substantives and adjectives, plus a few foreign prefixes like anti-, pro-, pseudo- etc. But ESTMORF treats 70 frequent initial components of words as Estonian native prefixes which can prefix a substantive, adjective, adverb or verb. In addition, there are 30 foreign prefixes which can prefix a substantive, adjective or verb.

In forming the lists of prefixes, ESTMORF has taken the following approach, based on purely formal criteria. A component should be listed as a prefix in its own rights if

1. The component cannot function as a word on its own, or has a clearly different meaning in compounds (e.g. ala 'area' meaning 'sub-' in compounds).
2. It is not trivial to see how the component was formed from some stem.
3. The component can be used freely to form new words.
4. It is frequent enough.

ESTMORF is fairly strict and instead of stating a doubtful rule, many derived words are kept in the lexicon. For example, the prefix nüüdis- 'contemporary' can be attached to substantives, e.g. nüüdisauto 'contemporary car', but not to adjectives, e.g. *nüüdispikk 'contemporary long'. However, a

few adjectives, like aegne ‘of the same time’, can also attach that prefix, e.g. nüüdisaegne ‘contemporary’, but as the number of such adjectives is limited it seems sensible to simply list them in the lexicon.

7. Analysing compound words

Rules and constraints for compound formation are related to two main characteristics:

1. The number of components in a word
2. The properties of the components themselves: e.g. is the component a stem or a derivational suffix; which word class does the stem belong to; which are the last letters of the stem etc.

In principle 8 different word patterns can participate in compound formation: stem, stem + inflectional affix, stem + derivational suffix, stem + derivational suffix + inflectional affix, prefix + stem, prefix + stem + inflectional affix, prefix + stem + derivational suffix, prefix + stem + derivational suffix + inflectional affix. Theoretically those patterns could combine in any manner, but in CELL texts the most popular combinations are:

Pattern	**% of all compounds**
stem + stem	70-75%
stem + stem + derivational suffix	5-10%
stem + stem + stem	5-10%
stem + inflectional affix + stem	1-5%
stem + inflectional affix + stem + derivational suffix	1-5%
stem + derivational suffix + stem	1-5%

Table II. The most frequent combinations for compounds

There are many restrictions which the components of every pattern must adhere to. The constraints for compounding are very much like those for derivation, involving the word class of the stem, the form of the stem (e.g. sometimes a stem of singular genitive may act as a component in a compound, but stem of a singular nominative cannot), and the ending letters of the stem. The restrictions of ESTMORF represent formal constraints only; no semantics is taken into account.

In addition, ESTMORF uses two lists of stems containing hundreds of tokens which tend to participate in compounds of different formation more often than others: more probable *initial components* and *final components* of compounds.

There may be several possibilities of splitting a compound word into components, e.g. lae+kaunistus ‘ornament of a ceiling’ and laeka+unistus ‘dream of a drawer’. ESTMORF finds only one possible splitting of a compound. The analysis of compounds is organised by choosing the sequence of subroutines and lists of stems in such a way that the output should be the most probable splitting, that is lae+kaunistus ‘ornament of a ceiling’. The primary guiding principle in doing so is minimising the number of components: prefer simplex readings to derived or compound ones and simpler compounds to more complex ones.

After several trials, we reached the following sequence of subroutines for parsing the structure of a compound or a derived word. This sequence gives the smallest error rate in determining the structures of words. The algorithm represents neither a left-to-right nor a right-to-left analysis, but rather a mixed one:

1. Is the string a simplex word?
2. Does the word have a structure *stem* + *derivational suffix* (or *stem* + *final component*)?
3. Does the word have a structure *prefix* + *stem* (or *initial component* + *stem*)?

4. Does the word have a structure *stem* + *stem*?
5. Does the word have a structure *stem* + *stem* + *derivational suffix* (or *stem* + *stem* + *final component*)?
6. Does the word have a structure *prefix* + *stem* + *derivational suffix* (or *initial component* + *stem* + *derivational suffix* or *prefix* + *stem* + *final component* or *initial component* + *stem* + *final component*)?
7. Does the word have a structure *stem* + *stem* + *stem*?
8. Does the word have a structure *stem* + *inflectional affix* + *stem*?
9. Does the word have a structure *stem* + *inflectional affix* + *stem* + *derivational suffix* (or *stem* + *inflectional affix* + *stem* + *final component*)?
10. Does the word have a structure *stem* + *derivational suffix* + *stem* (or *stem* + *derivational suffix* + *stem* + *derivational suffix* or *stem* + *derivational suffix* + *stem* + *final component*)?
11. Does the word have a structure *prefix* + *final component* (or *initial component* + *final component*)?
12. Does the word have a structure *prefix* + *compound word* (or *stem* + *compound word*)?

8. The ESTMORF lexicon

Creating a good lexicon for a morphological analyser is the most obvious benefit of using a corpus.

The ESTMORF lexicon contains 38,000 words. It is based on a machine-readable version of CMD [Viks 1992]. Because all the stem variants of a word are listed in the lexicon, the lexicon contains 67,000 entries. Comparing the ESTMORF lexicon with CMD we see that many words have been added:

1. About 1200 core vocabulary simplex words.
2. About 2500 compound words, the formation of which was too irregular or complex for describing algorithmically. These 2500 words represent the following word classes: 100 verbs, 870 adverbs, 150 numerals, 8 pronouns and 1300 substantives and adjectives
3. About 2700 proper names and 500 genitive attributes, among these about 70 names consisting of several words, e.g. New York.
4. About 200 abbreviations.
5. About 100 acronyms.

Thousands of words were deleted from CMD while forming the ESTMORF lexicon:

1. About 1800 obsolete or dialect simplex words.
2. About 2700 redundant derived words. (For some reason, CMD contains many productively derivable words).

9. Conclusion

When creating a morphological analyser and speller for Estonian, the aim was to create a fast program capable of analysing a raw text without artificial limitations. Lack of computational treatment of Estonian derivatives and compounds added to the difficulty of the task. After 5 years of development, we may say that we have achieved our goal.

The process of creating the analyser was an iterative one: first, a program was created, then it was checked on a corpus, the results were analysed and the program modified. Then the cycle repeated.

The methods we used during testing and corpus analysis were very simple. We did not go beyond frequency counts, percentages and simple comparisons of outputs.

Obviously, testing on a corpus had a tremendous impact on the lexicon. But it also enabled us to find an acceptable algorithm for analysing productive derivatives and compounds. In addition, testing on a corpus resulted in changes of the implementation of simplex word analysis.

10. Acknowledgements

ESTMORF would not have been possible without the electronic version of *A Concise Morphological Dictionary of Estonian* by Ü. Viks. Many crucial modules of ESTMORF were implemented by Tarmo Vaino. Viire Villandi selected and actually typed in various lists of proper names and used ESTMORF in its testing phases. Toomas Mattson, Ülle Viks, Heili Orav, Kadri Muischnek and Microsoft WPG used, tested and provided feedback on ESTMORF. The Department of General Linguistics of the University of Tartu provided all the corpora for developing and testing purposes. The author would like to thank Nancy Ide and the anonymous CHUM reviewers for their invaluable comments on the article.

11 . References

Brodda and Karlsson 1980 - Brodda, B. and Karlsson, F. "An Experiment with Automatic Morphological Analysis of Finnish." *Papers from the Institute of Linguistics. Publication 40*, Stockholm: University of Stockholm, 1980.

EKG 1995 -*Eesti Keele Grammatika 1*. (Grammar of the Estonian Language 1.); ed. M. Erelt, Tallinn: Eesti TA EKI, 1995

Francis et al. 1964 - N. W. Francis, H. Kucera, *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers.* Providence, R.I., 1964

Guidelines 1994 – *Guidelines for electronic Text Encoding and Interchange.* Ed. by Michael Sperberg-McQueen & Lou Burnard, Text encoding initiative. Chicago, Oxford. April 8, 1994

Hennoste et al. 1993 - T. Hennoste, K. Muischnek, H. Potter, T. Roosmaa 1993. "Tartu Ülikooli eesti kirjakeele korpus: ülevaade tehtust ja probleemidest." (The Tartu University Corpus of Estonian Literary Language: an overview of finished things and problems) *Keel ja Kirjandus*, 10 (1993). pp. 587-600.

Itogi 1983 - *VINITI Itogi nauki i tehniki. Serija informatika*, (VINITI Summaries of Science and Technology. Series of Informatics), Vol. 7. Moscow, 1985

Johansson et al. 1978 - S. Johansson, G. Leech, H. Goodluck, *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers.* Oslo, 1978

Karlsson 1992 - Karlsson, F. "SWETWOL: a Comprehensive Morphological Analyzer for Swedish." *Nordic Journal of Linguistics* 1, (1992), 1-45.

Kasik 1984 - R. Kasik. *Eesti keele tuletus›petus: õppevahend eesti filoloogia ja zurnalistikaosakonna üliõpilastele. 1. Substantiivituletus.* (Estonian derivation: a textbook for the students of the dept. of Estonian linguistics and journalism. 1. Derivation of substantives); TRÜ, Tartu 1984

Kasik 1992 - R. Kasik. *Eesti keele tuletusõpetus: õppevahend eesti filoloogia ja zurnalistikaosakonna üliõpilastele. 1. Adjektiivi- ja adverbituletus.* (Estonian derivation: a textbook for the students of the dept. of Estonian linguistics and journalism. 1. Derivation of adjectives and adverbs); TRÜ, Tartu 1992

Kask 1967 - Kask, A. "Liitsõnad ja liitmisviisid eesti keeles." (Compound Words and Ways of Compounding in Estonian). *Eesti keele grammatika 3.1*., Tartu, 1967

Koskenniemi 1983 - Koskenniemi, K. “Two-level Morphology: A General Computational Model for Wordform Recognition and Production.” *Publications of the Dept. Of General Linguistics, University of Helsinki*, 11 (1983)

Kull 1967 – Kull, R. *Liitnimisõnade kujunemine eesti kirjakeeles.* (Nominal compound development in Estonian literary language) Dissertation for candidate of philological sciences, ENSV TA KKI, Tallinn 1967

Proszeky and Tihanyi 1992 - Proszeky, G. and Tihanyi, L. “A fast Morphological Analyzer for Lemmatizing Agglutinative Languages.” Kiefer, F. G. Kiss and J. Pajzs (Szerk.) *Papers in Computational Lexicography. Complex-92*, Budapest: Linguistics Institute, HAS, 1992, pp. 265-278

Solak and Oflazer 1993 - A. Solak and K. Oflazer, “Design and Implementation of a Spelling Checker for Turkish.” *Literary and Linguistic Computing*, Vol. 8, No. 3, 1993

Sproat 1992 – R. Sproat, *Morphology and Computation.* The MIT Press, Cambridge, Mass.

Svartvik et al. 1980 - J. Svartvik, R Quirk, *A corpus of English conversation*. Lund, 1980

Valgma 1970 - J. Valgma, N. Remmel *Eesti Keele Grammatika*. (Grammar of the Estonian Language); Tallinn: Valgus, 1970

Viks 1992 - Ü. Viks *A Concise Morphological Dictionary of Estonian.* Tallinn: Institute of Estonian Language and Literature, 1992