# *Mees* 'man', *aeg* 'time' and other frequent words in the Corpus of Estonian Dialects

Liina Lindström[1], Eva Velsker, Ellen Niit, Karl Pajusalu

## 1. Introduction

In the present article we briefly describe the Corpus of Estonian Dialects (CED) compiled at the University of Tartu and give an overview of its current state. We will take a closer look at the most frequent case forms of the most frequent nouns in Estonian dialects and look for reasons why these forms are the most frequent ones. The present article is a follow-up to the overview article of the Corpus of Estonian Dialects published in 2001 (Lindström et al. 2001). In this 2001 article we compared the 100 most frequent word forms in three Estonian dialects and tried to find out the reasons for the frequent use of these word forms. In the present article we look at the three most frequent forms of the 50 most frequent nouns and the reasons for using them. Special attention is given to the possible cases of adverbialization and adposition formation, more broadly to the lexicalization and grammaticalization processes.

[1] Liina Lindström's work is supported by Estonian Science Foundation grant No. 7464.

The first part of the article briefly describes the Corpus of Estonian Dialects and discusses its state as of 20 August 2008. The second part of the article analyses the most frequent forms of the 50 most frequent nouns.

# 2. Overview of Estonian dialects and the Corpus of Estonian Dialects

## 2.1. Estonian dialects

What follows is a short overview of the distribution of Estonian dialect area as it is used in the dialect corpus. The present article and the dialect corpus adhere to the traditional division of dialects (see Pajusalu 2003). According to this division Estonian dialects are divided into three dialect groups. These dialect groups are further divided into different dialects. The following dialect groups and dialects are represented in the dialect corpus:

1) North Estonian dialect group: Mid, Eastern, Western, Insular dialects;
2) South Estonian dialect group: Võru, Mulgi, Tartu, Seto dialects;
3) North-Eastern Coastal dialect group: North-Eastern (Alutaguse), Coastal dialects.

Proceeding from Pajusalu et al. (2002) and differently from the traditional distribution, the North-Eastern Coastal dialect group is divided into two dialects: Coastal and North-Eastern (or Alutaguse in Pajusalu et al. 2002). Unlike in Pajusalu et al. (2002), in the dialect corpus the Seto dialect is taken as a separate dialect in the South Estonian dialect group. This reflects the desire to place more emphasis on this linguistically unique language in the southeast corner of South Estonia with its unique culture.

The dialects are further divided into parish dialects. In the second part of the article, the examples refer to dialects and the list of abbreviations of dialects is provided at the end of the article.

The following map shows the distribution of Estonian dialects which has been observed in compiling the corpus of dialects.

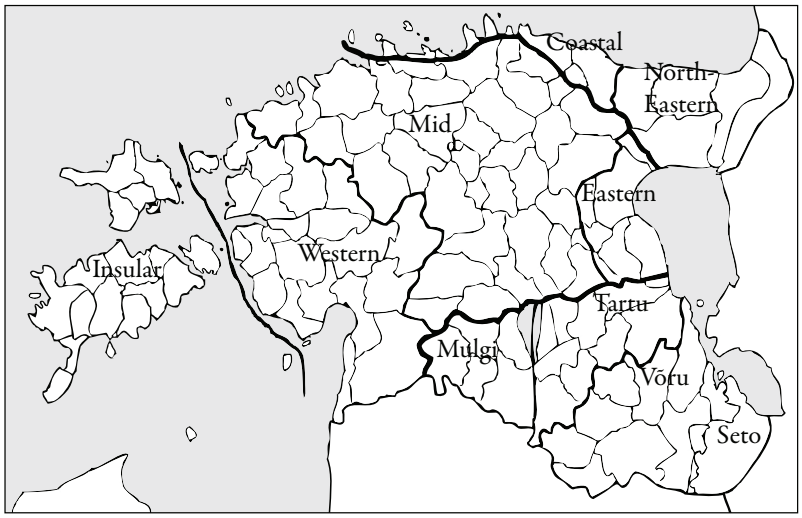


**Figure 1. Distribution of Estonian dialects.**

## 2.2. Corpus of Estonian Dialects (CED)

The Corpus of Estonian Dialects is an electronic data collection which includes authentic dialect texts from all Estonian dialects. Its main aim is to make accessible for the researchers dialect materials that are well-chosen and as accurately transcribed as possible. There is a comparable amount of material from every Estonian dialect in the corpus. To the first layer of CED which contains at least one million words of text (by the end of the year 2008) have been

selected the oldest available tape-recorded dialect texts. The dialect corpus should therefore represent relatively old dialect language.

The dialect corpus is compiled in cooperation of two institutions – the University of Tartu and the Institute of the Estonian Language. The materials used in the corpus come mainly from the Institute of the Estonian Language.

The dialect corpus consists of:
1) dialect recordings;
2) phonetically transcribed dialect texts;
3) dialect texts in simplified transcription;
4) morphologically tagged texts, which have been read into a MySQL database;
5) a database containing information about informants and recordings.

In the corpus, every phonetically transcribed text is accompanied by a recording, a file in simplified transcription and a description; more than half of the texts are also accompanied by a morphologically tagged file.

### 2.2.1. Dialect recordings

The corpus is based on dialect recordings which have mainly been made in the 1960s and 1970s. The first recordings date from 1938. The recordings are traditional dialect recordings where the interview is conducted at the home of the informant. Data on the years of the dialect recordings can be found in Table 1. The vast majority of the dialect recordings have been by now digitalised.

**Table 1. The years when the dialect recordings in the corpus were made (as of 20 August 2008).**

| Year of recording | Number of recordings |
|---|---|
| 1938 | 5 |
| 1957–1959 | 27 |
| 1960–1969 | 109 |
| 1970–1979 | 72 |
| 1980–1986 | 13 |
| 1991–1993 | 2 |
| unknown | 1 |
| Total | 229 |

### 2.2.2. Phonetically transcribed dialect texts

The dialect texts in Finno-Ugric phonetic transcription constitute one of the main parts of the corpus. The aim has been to transcribe the texts as accurately as possible; the phenomena accompanying spontaneous speech (e.g. the discourse particles, corrections, repetitions, etc.) have been added to the text, which traditionally have not been considered important in dialect research. The text of the interviewer has been transcribed as well.

The phonetically transcribed texts can only be opened with MS Word and in order to use them, one has to install the special fonts beforehand. The fonts were created by Esko Oja.

As of 20 August 2008 there are 964,398 words of phonetically transcribed text in the corpus. Table 2 gives the data according to the dialects. The proportion of texts from different dialect groups in the corpus is given in Figure 2.

**Table 2. Phonetically transcribed words in CED (as of 20 August 2008).**

| Dialect | Words | Proportion in the corpus |
|---|---|---|
| Eastern | 69,240 | 7.2% |
| Mid | 122,815 | 12.0% |
| Western | 146,605 | 15.2% |
| Insular | 195,971 | 20.3% |
| North-Eastern | 43,852 | 4.5% |
| Coastal | 55,246 | 5.7% |
| Mulgi | 52,468 | 5.4% |
| Tartu | 67,682 | 7.0% |
| Võru | 93,304 | 9.7% |
| Seto | 100,711 | 10.4% |
| Linguistic enclaves (in Latvia) | 16,504 | 1.7% |
| Total | 964,398 | |

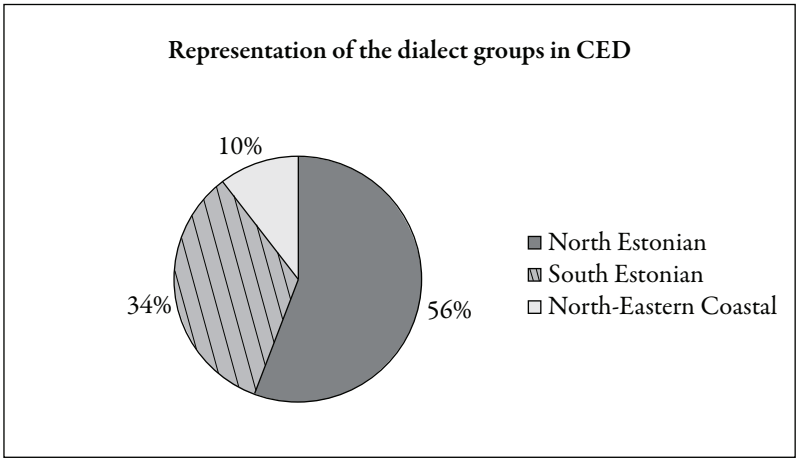


**Figure 2. Distribution of texts from different dialect groups in CED.**

### 2.2.3. Dialect texts in simplified transcription

All of the phonetically transcribed texts have been transported one-to-one into the simplified transcription (.txt), which enables the use of these texts with every program and to conduct primary analyses. There are some differences in the simplified transcriptions compared to the literary language; for example, the geminate voiceless consonants in quantity 2 words have been written with two letters (*kada**kk**ad*) and the acute accent inserted before the word helps to differentiate between quantity 2 and quantity 3 (in case of quantity 3, e.g. *'kas'si* 'cat' singular partitive). In case of a (indistinguishable) duration between quantity 2 and 3, the symbol * has been used (mainly in the North-Eastern Coastal dialect group, e.g. **kassi*). See Lindström (2005) for more details on simplified transcriptions. Also the palatalisation is marked by the apostrophe (*kas'si*).

In addition, the comments, the text of the informant(s) and the interviewer have been annotated in the simplified transcriptions.

### 2.2.4. Morphology database

Texts in the simplified transcription are morphologically tagged and read into the database. The texts were tagged with the help of the program Mark. The tagged texts are in XML format and they have been read into a MySQL database which can be used via the Internet. The database can be found at the following website: www.murre.ut.ee. At the moment, two search engines are in use; the first one (search.php) enables a more detailed analysis together with the context in which the word or form looked up appears, and the other one (search2.php) adds the statistics. In order to use the database, one needs to know the abbreviations and categories (word classes, morphological markers tagged, etc.) used in tagging. For more details on the abbreviations and categories see http://www.murre.ut.ee/otsing.html and Lindström et al. (2006), Lindström (2005).

As of 20 August 2008, there are 452,000 tagged tokens in CED. By the end of 2008 there should be 0.5 million tagged tokens.

For every word the following fields have been tagged:

**SNE**: the original form of the token as it occurs in the text, e.g. *t'ʒ'ibõrdõl'l'i* 'fidget' (past sg 3), *'vaesõq* 'poor' (pl nominative), *sääl* 'there'.

**MSN**: the keyword (lemma) in the literary language form (literary language spelling has been used and the vowel harmony has been lost), e.g. *tsiberdelema* 'to fidget'. If there is a word in the literary language with the same stem and meaning, the literary language word has been given as the keyword, e.g. *vaene* 'poor', *seal* 'there'.

**TAH**: meaning if it differs from the literary language. This field is not filled for every word; it is only marked, when the meaning of the word is different in the literary language or when there is no equivalent word in the literary language, e.g. *tsiberdelema* 'siplema' ('to fidget').

**FRA**: phrase, tagged for phrasal units, e.g. phrasal verbs (e.g. ***ära*** *ostma* 'buy'; *ära* 'away' is here the perfective marker).

**SLK**: word class. In the dialect corpus the main aim of determining the word classes has been to use the sort of classification which would be, on the one hand, understandable and detailed enough for researchers working with dialect languages, and on the other hand, clear-cut enough for those who do the morphological tagging.

Words have been divided into 26 word classes according to their morphological inflections, syntactic characteristics and semantics. This classification is based on the system of word classes presented in Estonian grammars (EKG I: 14–41); however, we distinguish more subclasses. Since the language in the dialect corpus is spoken, we also tagged some other phenomena related more to the spoken language use (discourse particles, communicative words; see Hennoste 2002). For more details on the issue of word classes in the dialect corpus see Lindström et al. (2006).

**MRF**: morphological information. Morphological information has been added to inflected words (nouns, verbs, pronouns,

adjectives, etc.). This information normally includes the form description consisting of a number of abbreviations, e.g. for the token *t'ś'ibõrdõl'l'i* the information presented in Table 3 is given in the database.

**KHK**: the parish where the word comes from. Abbreviations in capital letters have been used (VAS=Vastseliina etc., see, for example, http://www.eki.ee/murded/). In addition to the data on Estonian dialects, Votic texts have also been tagged to a certain extent (the corresponding parish abbreviations are IVA=Eastern-Votic and LVA=Western-Votic).

**Table 3. Description of the token *ťśibõrdõlľi* in the morphological database.**

| SNE | MSN | TAH | FRA | SLK | MRF | KHK |
|---|---|---|---|---|---|---|
| ťśibõrdõlľi | tsiber-delema | siplema | | V | ps ind ipf pl 3 | HAR |

### 2.2.5. Information about informants and recordings

An Access database has been compiled for every dialect material used. This database includes as much information as possible about the informants, the recordings and the transcriptions. As these are old recordings, some of them have certain information missing and it has not been possible to obtain it afterwards. The following information is given for every recording:

- the number of the tape (digital recording) in the corpus, the dialect, the subdivision of the dialect, village;
- name, age, date of birth, and other obtainable personal data (education, descent, parents, etc.) about the informant;
- date of the recording, names of the people who did the recording, the background of the tape (tape number in the sound library of the Institute of the Estonian Language or the University of Tartu);

- transcription number in the Institute of the Estonian Language or the University of Tartu, the person who transcribed, the person who checked the transcription and inserted the text;
- amount of the text part (in words) included in the corpus, the amount of words tagged.

## 3. Most frequent nouns in the dialect corpus

In this section, we analyse the 100 most frequent nouns in the CED and compare them to similar data taken from the literary language corpus. We look at what are the most frequent forms of the 50 most frequent nouns. We are interested in whether the frequent use of certain word forms shows that this word form has separated from the noun paradigm and started, so to speak, its own life, i.e. it has become an adverb or an adposition. In order to answer this question we look at whether the same word form has been interpreted differently by the people who do the morphological tagging and who have tagged it as some other (uninflected) word class. We furthermore try to find out the reasons for this difference in tagging.

In 2001 we conducted a similar study based on tokens; we looked at the most frequent tokens in three dialects (Lindström et al. 2001). The present study differs from the previous one in many respects:

1) in the present study we use the tagged text, i.e. the lemmas, as the basis. Thus, the small differences in the phonetic form of the word no longer influence the results;

2) we look at the three most frequent forms of the most frequent nouns in order to determine the form in which certain words are used most naturally. When a certain meaning has become fixed

in a certain form, it may be the reason why this form has become detached from the noun paradigm; it may indicate either lexicalization or grammaticalization;

3) we do not concentrate on the differences between dialects, but rather compare the data from the dialect corpus to the data found in the frequency dictionary based on the literary language corpus (Kaalep, Muischnek 2002). As a result we can outline the differences between dialect and literary language and the vocabulary characteristic of the traditional Estonian rural culture.

Following is a list of the 100 most frequent nouns in the dialect corpus as of 1 June 2008. At the time of the query the database contained 370,586 tagged words from Estonian dialects; about 66,137 of them (that is about 18%) were nouns.

**Table 4. The most frequent nouns in CED. For each word, the position in the frequency list and the number of occurences is given.**

| | | | | | |
|---|---|---|---|---|---|
| 1. | *mees* 'man, husband' | 968 | 51. | *nimi* 'name' | 213 |
| 2. | *aeg* 'time' | 955 | 52. | *vili* 'grain' | 211 |
| 3. | *inimene* 'human' | 739 | 53. | *kartul* 'potato' | 210 |
| 4. | *laps* 'child' | 723 | 54. | *lammas* 'sheep' | 207 |
| 5. | *isa* 'father' | 596 | 55. | *kirik* 'church' | 200 |
| 6. | *ema* 'mother' | 571 | 56. | *tuba* 'room' | 199 |
| 7. | *aasta* 'year' | 569 | 57. | *vend* 'brother' | 196 |
| 8. | *maa* 'land, ground' | 568 | 58. | *loom* 'animal' | 190 |
| 9. | *hobune* 'horse' | 564 | 59. | *tuul* 'wind' | 189 |
| 10. | *päev* 'day' | 532 | 60. | *talv* 'winter' | 185 |
| 11. | *vesi* 'water' | 531 | 61. | *auk* 'whole' | 184 |
| 12. | *asi* 'thing' | 531 | 62. | *tütar* 'daughter' | 183 |
| 13. | *töö* 'work' | 526 | 63. | *piim* 'milk' | 182 |
| 14. | *kord* 'time, turn' | 482 | 64. | *selg* 'back' | 176 |
| 15. | *mõis* 'manor' | 473 | 65. | *pulm* 'wedding' | 176 |
| 16. | *talu* 'farm' | 456 | 66. | *tuli* 'fire' | 165 |
| 17. | *naine* 'woman, wife' | 454 | 67. | *saun* 'sauna' | 175 |
| 18. | *puu* 'tree' | 425 | 68. | *jää* 'ice' | 174 |

| | | | | | |
|---|---|---|---|---|---|
| 19. | *ots* 'tip, end' | 413 | 69. | *põld* 'field' | 172 |
| 20. | *koht* 'place' | 397 | 70. | *linn* 'town' | 171 |
| 21. | *pea* 'head' | 379 | 71. | *suvi* 'summer' | 169 |
| 22. | *võrk* 'net' | 371 | 72. | *silm* 'eye' | 169 |
| 23. | *mets* 'forest' | 366 | 73. | *pruut* 'bride' | 169 |
| 24. | *küla* 'village' | 359 | 74. | *peremees* 'master' | 168 |
| 25. | *kala* 'fish' | 345 | 75. | *siga* 'pig' | 168 |
| 26. | *poeg* 'son' | 326 | 76. | *elu* 'life' | 162 |
| 27. | *raha* 'money' | 325 | 77. | *pere* 'family' | 156 |
| 28. | *käsi* 'hand' | 317 | 78. | *nädal* 'week' | 159 |
| 29. | *jalg* 'foot' | 306 | 79. | *kodu* 'home' | 159 |
| 30. | *tükk* 'piece' | 297 | 80. | *liha* 'meat' | 142 |
| 31. | *leib* 'bread' | 297 | 81. | *masin* 'machine' | 153 |
| 32. | *meri* 'sea' | 295 | 82. | *viin* 'vodka' | 142 |
| 33. | *paat* 'boat' | 293 | 83. | *kott* 'bag' | 141 |
| 34. | *lina* 'sheet' | 292 | 84. | *hein* 'hay' | 141 |
| 35. | *ahi* 'stove' | 284 | 85. | *uss* 'snake' | 141 |
| 36. | *poiss* 'boy' | 276 | 86. | *nahk* 'skin' | 140 |
| 37. | *kivi* 'stone' | 273 | 87. | *põhi* 'bottom' | 136 |
| 38. | *õhtu* 'evening' | 266 | 88. | *muna* 'egg' | 102 |
| 39. | *kangas* 'fabric' | 266 | 89. | *rand* 'beach' | 135 |
| 40. | *maja* 'house' | 254 | 90. | *ilm* 'weather' | 133 |
| 41. | *tüdruk* 'girl' | 252 | 91. | *jõgi* 'river' | 130 |
| 42. | *laud* 'table' | 252 | 92. | *rukis* 'rye' | 129 |
| 43. | *kari* 'cattle' | 245 | 93. | *laev* 'boat' | 128 |
| 44. | *tee* 'road, way' | 244 | 94. | *kool* 'school' | 126 |
| 45. | *lehm* 'cow' | 239 | 95. | *kell* 'watch, clock' | 126 |
| 46. | *lõng* 'yarn' | 235 | 96. | *rahvas* 'people, nation' | 123 |
| 47. | *riie* 'cloth' | 226 | 97. | *õlu* 'beer' | 123 |
| 48. | *hommik* 'morning' | 222 | 98. | *rohi* 'grass' | 118 |
| 49. | *meel* 'sense' | 216 | 99. | *pulk* "stick' | 116 |
| 50. | *rehi* 'threshing barn' | 214 | 100. | *tanu* 'coif' | 116 |

In the literary language corpus of fiction and newspaper texts the most frequent nouns are completely different. Following is a list of lemmas taken from among the 10,000 most frequent lemmas in the net version of Kaalep and Muischnek (2002) which carry the word class label noun (S)[2]. As they have not indicated the word class for the corresponding word in every case, the frequency list also contains those whose high frequency of use is due to their use as pronouns, conjunctions or as another word class, but the same word shape can also be used as a noun. The following words in the list are like this: *mina* 'I', *oma* 'own', *sina* 'you', *kõik* 'all', *iga* 'every' (typically pronouns and in the dialect corpus they are used as pronouns), *või* 'or' (usually a conjunction), *siin* 'here' (usually a pro-adverb), etc. When tagging the texts in the dialect corpus, it has been decided in every concrete context to which word class the word shape belongs.

Following is a list of the 100 most frequent nouns according to their frequency taken from the 1990s corpus of literary language: *mina* 'I', *oma* 'own', *sina* 'you', *või* 'or; butter', *aasta* 'year', *kõik* 'all', *mees* 'man, husband', *aeg* 'time', *inimene* 'human', *sõna* 'word', *kord* 'time, turn', *naine* 'woman, wife', *pool* 'side; half', *käsi* 'hand', *päev* 'day', *kroon* 'crown, Estonian kroon', *laps* 'child', *iga* 'every', *asi* 'thing', *pea* 'head', *riik* 'state', *silm* 'eye', *siin* 'here', *töö* 'work', *elu* 'life', *hea* 'good', *raha* 'money', *enne* 'before; omen', *linn* 'town', *kogu* 'whole', *tee* 'road, way', *mõte* 'idea', *ema* 'mother', *maja* 'house', *kodu* 'home', *koht* 'place', *vana* 'old', *juht* 'leader', *maa* 'land, ground', *osa* 'part', *valitsus* 'government', *koos* 'together', *nägu* 'face', *poiss* 'boy', *maailm* 'world', *keel* 'language', *auto* 'car', *võimalus* 'chance', *uks* 'door', *vene* 'Russian', *kuu* 'month, moon', *isa* 'father', *noor* 'young', *küsimus* 'question', *hääl* 'voice', *aru* 'reason', *tüdruk* 'girl', *nädal* 'week', *vesi* 'water', *viis* 'five', *lõpp* 'end', *rahvas* 'people,

[2] The 10,000 most frequent lemmas can be found at the following webpage: http://www.cl.ut.ee/ressursid/sagedused/tabel1.txt.

nation', *nimi* 'name', *kool* 'school', *hetk* 'moment', *jutt* 'story, line', *paar* 'pair', *kell* 'watch, clock', *meel* 'sense', *jalg* 'foot', *pilk* 'glance', *kaasa* 'spouse; with', *seadus* 'law', *tänav* 'street', *õhtu* 'evening', *liiga* 'league', *president* 'president', *firma* 'firm', *protsent* 'per centage', *laud* 'table', *tuba* 'room', *valge* 'white', *pank* 'pank', *riigikogu* 'parliament', *õigus* 'right', *liit* 'union', *esimees* 'chairman', *tund* 'hour', *selg* 'back', *liige* 'member', *otsus* 'decision', *hommik* 'morning', *raamat* 'book', *lugu* 'story', *algus* 'beginning', *hind* 'price', *olukord* 'situation', *põhjus* 'reason', *aken* 'window', *terve* 'whole, healthy', *võõras* 'stranger', *eestlane* 'Estonian', *nõukogu* 'council'.

As can be seen, most of the frequent vocabulary in modern Estonian is the same as in the traditional dialect recordings, where the speakers were mostly born in the 19th century. Despite the times and the changed world, the most frequently used words are still *mees* 'man, husband', *inimene* 'human', *naine* 'woman, wife', *laps* 'child', *ema* 'mother', *isa* 'father', *maja* 'house', *kodu* 'home', *töö* 'work', *aasta* 'year', *pea* 'head', *käsi* 'hand', *silm* 'eye', *jalg* 'foot', *maa* 'land, ground', *asi* 'thing', *rahvas* 'people, nation', *linn* 'town', etc.

The words that compared to the literary language corpus are different in the dialect corpus are mostly related to the traditional rural culture; this group of words includes the vocabulary related to agriculture and handicraft, also to the life and times of the 19th century when people lived in farms and worked in manors. This is the reason why the words on top of the frequent words list in the dialect corpus are nouns denoting different tools like *hobune* 'horse', *paat* 'boat', *võrk* 'net', institutions *talu* 'farm', *mõis* 'manor', *kirik* 'church', *küla* 'village', traditional farm buildings *rehi* 'threshing barn' and *saun* 'sauna', handicraft *kangas* 'fabric', *riie* 'cloth', lõng 'yarn', domestic animals *lammas* 'sheep', *lehm* 'cow', *siga* 'pig', *kari* 'cattle', crop *kartul* 'potato', *lina* 'linen', *vili* 'grain', *rukis* 'rye', agricultural products *piim* 'milk', *liha* 'meat', *muna* 'egg'. Other words refer to the customs and include *pulm* 'marriage', *pruut* 'bride', *tanu* 'coif'. Landscape objects are also more often referred to in the dialect corpus than in the modern literary language; in the top 100

are words like *mets* 'forest', *meri* 'sea', *puu* 'tree', *kivi* 'stone', *põld* 'field', *rand* 'beach', *jõgi* 'river'. The frequent use of words referring to such traditional rural culture is due not only to the fact that the speakers were mostly born in the 19th century, but also because the interviewers have especially asked the speakers to talk about old things, handicraft, old traditions, etc. (see also Lindström 2001).

## 3.1. The role of frequency in grammaticalization and lexicalization

There are, however, a large number of words among the most frequent nouns in the dialect corpus, in which case it cannot be understood why these words are on top of the frequency list (e.g. *ots* 'tip, end'). It seems that certain fixed expressions have also been labelled as nouns, which could actually be seen as stages in the process of grammaticalization or lexicalization.

It is known from the grammaticalization theory that during the process of grammaticalization the meaning of a word form becomes more general, more abstract and that the use of the form extends (Heine 2005). On the one hand, when meaning becomes more abstract and it is used in wider contexts, the word form becomes more frequent – the more the form has grammaticalized, the more frequent it is (Hopper and Traugott 2003: 113). On the other hand, as pointed out by Joan Bybee, frequency itself also affects the weakening of the meaning. Bybee refers to this process as habituation – „the process by which an organism ceases to respond at the same level to a repeated stimulus" (Bybee 2005: 604). Frequent repetition makes the word form also phonologically reduced. Bybee considers the role of frequent repetition so important that she defines grammaticalization as a process when „a frequently used sequence of words or morphemes becomes automated as a single processing unit" (Bybee 2005: 603).

Nevertheless, frequency is not always related to the grammaticalization process, but the frequent use itself may cause the separation of the word form from the rest of the paradigm and its treatment as an independent unit in the lexicon of a speaker (Bybee 2007: 13). One such example is the illative form *koju* of the word *kodu* 'home', which is formed irregularly and is thus autonomous in the lexicon of a speaker, separate from the rest of the forms of the word *kodu*. Thus, the frequent use of certain word forms may be instead related to the lexicalization process. Frequency plays, therefore, an important role both in grammaticalization as well as in lexicalization process. In the former case the new unit becomes grammatical (e.g. becomes an adposition), in the latter case the word form detaches itself from the general paradigm and becomes autonomous (adverbialization).

Next we will take a closer look at the three most frequent forms of the 50 most frequent words in the dialect corpus and try to find out why it is these forms that are used most frequently. Should any word form show signs of becoming either an adverb or adposition, we study whether the same word form has been also tagged differently from nouns, i.e. as an adverb or an adposition. We exclude from our analysis the idea that the differences in the tags of the same word form are the mistakes made by the people who have tagged the texts (although it could very well be the case), but that these differences give us information about the fuzzy areas between different word classes.

## 3.2. The most frequent word forms of the 50 most frequent words in the dialect corpus

Table 5 gives the three most frequent forms of the 50 most frequent nouns, the number of the forms found in the corpus and the proportion of each form in the total number of uses of this word (%).

Certain semantic groups emerge from this table, which have a very similar frequency of the three forms. Such, for example, are PEOPLE, OBJECTS, BODY PARTS, PLACE, TIME, OTHER. We will next look at the results presented in Table 4 according to these semantic groups.

**3.2.1.** PEOPLE, LIVING BEINGS. Words referring to people are on top of the frequency list of the dialect corpus: *mees* 'man, husband', *inimene* 'human', *laps* 'child', *naine* 'woman, wife', *isa* 'father', *ema* 'mother', *poeg* 'son', *poiss* 'boy', *tütar* 'daughter'. What they all have in common is that the nominative is the most frequent case. In all probability the words referring to people typically act as agents and are generally used as subjects; this is the reason why the nominative plays an important role. With some words, the nominative plural is the most frequent case, e.g. *inimene* 'human' and *laps* 'child' (*inimesed* 'humans', *lapsed* 'children'); with the rest of the words, the nominative plural is the second most frequent case after the nominative singular. The third most frequent form is the genitive singular (sg g, e.g. *mehe* 'man's, husband's') or partitive singular (sg p, e.g. *poega* 'son'). Both are typically object cases. (In Estonian, the object can be either in the nominative, genitive, or partitive; see Erelt 2003: 96). The genitive is furthermore used in the possessive construction and in adposition constructions. Thus, the frequent use of the so-called grammatical cases (nominative, genitive, and partitive) is completely predictable. These are the most frequent cases in Estonian in general, both in written (Valge 1970) and spoken language (Hennoste 2004).

**Table 5. The most frequent forms of the 50 most frequent nouns.**

| | | 1. | | | 2. | | | 3. | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Keyword | Total | form | N | % | form | N | % | form | N | % |
| ***mees*** **'man, husband'** | 968 | sg n | 403 | 41,6% | pl n | 193 | 19,9% | sg g | 89 | 9,2% |
| ***aeg*** **'time'** | 955 | sg ad | 332 | 34,8% | sg n | 211 | 22,1% | sg p | 199 | 20,8% |
| ***inimene*** **'human'** | 739 | pl n | 247 | 33,4% | sg n | 207 | 28,0% | sg p | 79 | 10,7% |
| ***laps*** **'child'** | 723 | pl n | 182 | 25,2% | sg n | 160 | 22,1% | sg p | 79 | 10,9% |
| ***isa*** **'father'** | 596 | sg n | 409 | 68,6% | sg g | 92 | 15,4% | sg ad | 37 | 6,2% |
| ***ema*** **'mother'** | 571 | sg n | 400 | 70,1% | sg g | 68 | 11,9% | sg ad | 40 | 7,0% |
| ***aasta*** **'year'** | 569 | sg p | 277 | 48,7% | sg g | 100 | 17,6% | sg n | 88 | 15,5% |
| ***maa*** **'land, ground'** | 568 | sg g | 130 | 22,9% | sg p | 113 | 19,9% | sg n | 93 | 16,4% |
| ***hobune*** **'horse'** | 564 | sg n | 117 | 20,7% | sg g | 104 | 18,4% | pl n | 77 | 13,7% |
| ***päev*** **'day'** | 532 | sg p | 133 | 25,0% | sg n | 125 | 23,5% | sg g | 106 | 19,9% |
| ***vesi*** **'water'** | 531 | sg n | 172 | 32,4% | sg g | 144 | 27,1% | sg p | 109 | 20,5% |
| ***asi*** **'thing'** | 531 | sg n | 249 | 46,9% | sg p | 107 | 20,2% | pl n | 64 | 12,1% |
| ***töö*** **'work'** | 526 | sg p | 203 | 38,6% | sg n | 90 | 17,1% | sg g | 48 | 9,1% |
| ***kord*** **'time, turn'** | 482 | sg n | 136 | 28,2% | sg p | 136 | 28,2% | sg g | 96 | 19,9% |
| ***mõis*** **'manor'** | 473 | sg g | 182 | 38,5% | sg in | 104 | 22,0% | sg n | 58 | 12,3% |
| ***talu*** **'farm'** | 456 | sg n | 125 | 27,4% | sg g | 99 | 21,7% | sg in | 51 | 11,2% |
| ***naine*** **'woman, wife'** | 454 | sg n | 134 | 29,5% | pl n | 124 | 27,3% | sg g | 53 | 11,7% |
| ***puu*** **'tree'** | 425 | sg g | 102 | 24,0% | sg n | 77 | 18,1% | sg el; pl n | 70 | 16,5% |
| ***ots*** **'end, tip'** | 413 | sg n | 102 | 24,7% | sg in | 84 | 20,3% | sg g | 56 | 13,6% |
| ***koht*** **'place'** | 397 | sg in | 93 | 23,4% | sg g | 70 | 17,6% | sg n | 59 | 14,9% |
| ***pea*** **'head'** | 379 | sg n | 76 | 20,1% | sg g | 75 | 19,8% | sg ill | 73 | 19,3% |
| ***võrk*** **'net'** | 371 | pl n | 150 | 40,4% | sg g | 56 | 15,1% | sg p | 47 | 12,7% |
| ***mets*** **'forest'** | 366 | sg ill | 115 | 31,4% | sg in | 83 | 22,7% | sg g | 60 | 16,4% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ***küla*** **'village'** | 359 | sg g | 107 | 29,8% | sg in | 90 | 25,1% | sg n | 51 | 14,2% |
| ***kala*** **'fish'** | 345 | pl p | 72 | 20,9% | sg p | 67 | 19,4% | sg n | 66 | 19,1% |
| ***poeg*** **'son'** | 326 | sg n | 134 | 41,1% | pl n | 58 | 17,8% | sg p | 38 | 11,7% |
| ***raha*** **'money'** | 325 | sg p | 195 | 60,0% | sg g | 51 | 15,7% | sg n | 49 | 15,1% |
| ***käsi*** **'hand'** | 317 | sg ill | 56 | 17,7% | pl n | 38 | 12,0% | sg g | 37 | 11,7% |
| ***jalg*** **'leg'** | 306 | pl n | 90 | 29,4% | sg g | 43 | 14,1% | sg n; sg p | 28 | 9,2% |
| ***tükk*** **'piece'** | 297 | sg p | 168 | 56,6% | sg n | 41 | 13,8% | pl n | 35 | 11,8% |
| ***leib*** **'bread'** | 297 | sg p | 121 | 40,7% | sg n | 69 | 23,2% | sg g | 50 | 16,8% |
| ***meri*** **'sea'** | 295 | sg g | 96 | 32,5% | sg in | 48 | 16,3% | sg ill | 45 | 15,3% |
| ***paat*** **'boat'** | 293 | sg g | 76 | 25,9% | sg n | 69 | 23,5% | sg ill; sg com | 30 | 10,2% |
| ***lina*** **'flax, linen'** | 292 | sg g | 64 | 21,9% | pl n | 60 | 20,5% | pl p | 58 | 19,9% |
| ***ahi*** **'stove'** | 284 | sg ill | 88 | 31,0% | sg g | 68 | 23,9% | sg n | 53 | 18,7% |
| ***poiss*** **'boy'** | 276 | sg n | 122 | 44,2% | pl n | 93 | 33,7% | sg g | 15 | 5,4% |
| ***kivi*** **'stone'** | 273 | pl n | 75 | 27,5% | sg n | 48 | 17,6% | sg g | 49 | 17,9% |
| ***õhtu*** **'evening'** | 266 | sg g | 140 | 52,6% | sg n | 77 | 28,9% | sg ad | 14 | 5,3% |
| ***kangas*** **'fabric'** | 266 | sg n | 81 | 30,5% | sg g | 59 | 22,2% | sg p | 49 | 18,4% |
| ***maja*** **'house'** | 254 | sg n | 78 | 30,7% | sg g | 54 | 21,3% | sg in | 35 | 13,8% |
| ***tüdruk*** **'girl'** | 252 | sg n | 97 | 38,5% | pl n | 75 | 29,8% | sg g, sg p | 19 | 7,5% |
| ***laud*** **'table'** | 252 | sg g | 107 | 42,5% | sg n | 40 | 15,9% | pl n | 26 | 10,3% |
| ***kari*** **'cattle'** | 245 | sg in | 102 | 41,6% | sg n | 46 | 18,8% | sg g | 32 | 13,1% |
| ***tee*** **'way, road'** | 244 | sg g | 109 | 44,7% | sg n | 63 | 25,8% | sg p | 33 | 13,5% |
| ***lehm*** **'cow'** | 239 | sg p | 65 | 27,2% | pl n | 46 | 19,2% | sg n | 38 | 15,9% |
| ***lõng*** **'yarn'** | 235 | sg p | 60 | 25,5% | pl n | 44 | 18,7% | sg n | 41 | 17,4% |
| ***riie*** **'cloth'** | 226 | pl g | 47 | 20,8% | pl n | 45 | 19,9% | pl p | 39 | 17,3% |
| ***hommik*** **'morning'** | 222 | sg g | 169 | 76,1% | sg ad | 17 | 7,7% | | | 0,0% |
| ***meel*** **'sense, mind'** | 216 | sg in | 117 | 54,2% | sg ill | 42 | 19,4% | sg n | 24 | 11,1% |
| ***rehi*** **'threshing barn'** | 214 | sg g | 79 | 36,9% | sg p | 44 | 20,6% | sg n | 27 | 12,6% |

Two words stand apart from the rest, *isa* 'father' and *ema* 'mother', which refer to unique people from the viewpoint of the speaker. This is probably the reason why the nominative plural is not among the three most frequent forms of these two words. The proportion of nominative singular for these words is much bigger than in case of other words referring to people; for *isa* 'father' the percentage is 68 and for *ema* 'mother' even 70. In case of these two words, the second most frequent form is the genitive singular, the third one is the adessive singular. Both, especially the adessive singular, indicate the frequent use of the word form in the possessive construction (example 1).

(1) *mu **emal** oli üks ristiga rubel* (Mus)
'My **mother** had a rouble with a cross'

The word *hobune* 'horse' is used in a similar way to words referring to people, but in this case the nominative singular does not dominate as much as with *ema* 'mother' or *isa* 'father'. The three most frequent cases only make up about 53% of the total number of uses; thus, other cases are also used a lot. The most frequent form of the word *lehm* 'cow' is the partitive singular; this is probably due to the use of this form in quantity phrases (it was important that there was more than one cow in the farm), e.g.:

(2) *sin maeas õli *seitse *lehma* (Lüg)
'There were seven **cows** in this household'

Therefore, it can be said that horse as an important working animal is seen more as an individual in the dialect texts than, for example, a cow, and if we compare the data on frequency, even more than a child.

The word *kala* 'fish' in the group of words referring to living beings behaves completely in its own way; the most frequent form is the partitive plural (example 3). This can be probably explained

by the abundance of nautical texts – fish were something to be caught and they were caught in quantities, not separately. The word *kalu* 'fish' (partitive plural) is mainly used in sentences as an object (example 3).

(3) *ja *püüdasin **kalu*** (Lüg)
'and I caught **fish**'

Words referring to people and living beings are thus mainly used as the core arguments of a sentence, as a subject or object; others (especially *ema* 'mother' and *isa* 'father') are often used also as the adessive possessor or as adverbial marking the experiencer.

**3.2.2.** OBJECTS. *Asi* 'thing', *puu* 'tree', *võrk* 'net', *raha* 'money', *tükk* 'piece', *leib* 'bread', *ahi* 'stove', *kivi* 'stone', *laud* 'table' are the most frequent inanimate objects. Objects do not form as clear group as do people. Here, a number of words have their own specific patterns of use.

The word *asi* 'thing' has a very general meaning and it can often mean something like 'a problem, matter' (example 4) and it is used to refer to some kind of general situation (example 5). Nevertheless, in dialect texts it is more frequently used in the nominative singular case (example 5); but the partitive singular is also frequent (example 6).

(4) ** tegima ***asja** *selgest isaga* (Lüg)
'dad and I resolved the **matter**'
(5) *sie **as'i** mull seisab viel mieles* (Avi)
'I still remember this **thing**'
(6) *miss te *keikki *tühja *as'ja sis *küs'sitte* (Jõe)
'why do you ask all these unimportant **things**'

Words *tükk* 'piece', *leib* 'bread' and *raha* 'money' are similar because the most frequent form for all is the partitive singular. With the words *raha* and *tükk* the partitive is more frequent; they typically modify quantity phrases (example 7). The word *raha* typically functions as an object in the partitive case (denotes indefinite quantity, example 8).

(7) *ma 'saijõ sada rubla **raha*** (Kam)
'I got hundred rubles of **money**'
(8) *siss kui **raha** kor'jat'ti jälle* (Kam)
'when **money** was collected again'

The partitive use of the word *tükk* can be explained by grammaticalization. In Estonian, this word is often used as a pronoun in quantity phrases; it replaces countable objects (example 9):

(9) *`lehmi kaa yks kümme '**tük'k'ü** vähämbalt ol'l' sääl majan* (Har)
'this household had cows as well, at least ten' (lit. 'at least ten **pieces**')

It is characteristic of the word *puu* 'tree' that the third most frequent case form is elative (*puust* 'from the tree'), which indicates material (example 10):

(10) *suured pikkad udjad ol'lid **puust** udjad* (Hää)
'they were long poles, made **from wood**'

Thus other words referring to objects are also used as the core arguments of a sentence in the dialect corpus, especially as objects. Nevertheless, some words indicate certain levels of grammaticalization (for example, *tükk* 'piece' and possibly also *asi* 'thing').

**3.2.3.** BODY PARTS. Among the 50 most frequent body part words are *pea* 'head', *käsi* 'hand', *jalg* 'foot'; we may include here also *meel* 'sense, mind' with certain reservations. The grammaticalization of body part nouns is extremely well-known in the grammaticalization theory and it is a universal phenomenon (Heine 2005). In Estonian too, there are adpositions formed from the body part *pea* 'head': *peal* 'on (top of)' (*laua peal* 'on the table'), *peale* 'onto' (*laua peale* 'onto the table'; *peale vihma* 'after the rain'), *pealt* 'from (the top of)'; *käsi* 'hand': *käes* 'in hand' (*vihma käes* 'in [the hand of] the rain'), *kätte* 'to hand', *käest* 'from hand' (*vihma käest* 'from the [hand of the] rain'). Clearly grammaticalized cases have not been considered as nouns in the present frequency list because they are tagged as adpositions.

Let us look at, first of all, the paired body parts *käsi* 'hand' and *jalg* 'foot'. It is relatively predictable that the nominative plural is among the three most frequent forms; for the word *jalg* it is the most frequent form, for the word *käsi* it is the second most frequent. Surprisingly, the most frequent form of the word *käsi* is the illative singular. There are two reasons for this – illative is frequently used in the context of clothing (as in example 11), but also more generally in the context of keeping something in somebody's possession, while the contact with hands is clearly present (examples 12 and 13). This type of use has adverbialized to a certain extent, at least the people who have tagged the corpus have perceived it so, because some similar examples have been labelled as adverbs. The use of the illative form as an affixal adverb in particle verbs[3] has contributed to the adverbialization, e.g. *kätte saama* 'catch' (lit. 'get into hand', example 14). Majority of such cases have actually been tagged as affixal adverbs, some also as the illative forms of the noun. Since differentiating between affixal adverbs, adverbs and the illative form of nouns is not always easy, it is understandable why the illative form of the noun is used so frequently.

[3] Particle verbs are idiomatic or non-idiomatic verbs, which include particles referring to place, perfectivity, condition or modality. (Erelt 2003: 101)

(11) *kass ma aean *kinda **käde*** (Khn)
'do I put on my glove' (lit. 'do I put/drive the glove into the hand')
(12) *ja ˈjälle adra ˈ**kätte** ja kaks oost ˈette* (Vän)
'and again you grab the plough and put two horses in front' (lit. 'and again the plough into the hands and...')
(13) *võtta paar obusid ˈ**kätte** ja ja v- vea* (Muh)
'take a couple of horses in your **hand** and pull'
(14) *et sa ta **käde** saad* (Khk)
'that you catch him' (lit. 'get him into your **hands**')

The illative form is frequently used also with the nouns *pea* 'head' and *meel* 'mind, sense'; with the word *meel*, the only form more frequent than the illative is the inessive form. With the word *meel* the frequency of the illative derives from the frequent use of the particle verbs *meelde tulema* 'come to mind' (example 15), *meelde jääma* 'stick in one's mind' (example 16), and *meelde tuletama* 'bring sth back to sb's mind' (example 17). These again derive from the fact that dialect texts are predominantly recollections. The frequent use of the inessive form is due to the frequent use of the particle verbs *meeles olema* 'remember' (example 18) and *meeles seisma* 'stick in one's mind'. Expression verbs are idiomatic units, which include a noun in a certain fixed form (Erelt 2003: 101–102). In the dialect corpus such nouns have been labelled both as fully meaningful nouns as well as affixal adverbs. The changes in the meaning of the noun are apparent and it is clearly a unit which has become detached from the noun paradigm; here we are dealing with the lexicalization of a unit consisting of a verb and a noun.

(15) *p tule ˈ**meele** änam* (Mus)
'it doesn't come to my **mind** anymore'
(16) *poistel jäi ***miele** et* (Kuu)
'it stuck in the boys' **mind** that'

(17) *jah no tulettage muul ˈ**meeli** kedagi* (Mih)
'and help me **remember** somebody'
(18) *mull eij olõ mp ***meelen*** (Rõn)
'I no longer **remember**' (lit. 'I don't have it in my mind anymore')

The illative forms of the word *pea* 'head' generally denote the corresponding body part (examples 19 and 20) and this word is most often used in relation to clothing (example 19). These forms have a different morphology and have thus become separate from the general paradigm (the regular form should be *peasse*); this is why we may predict certain lexicalization in such cases. Cases where there has been a meaning transfer and the unit has become idiomatic have also been labelled as illative forms, e.g. as part of expression verbs *pähe hakkama* 'memorise; to go to the head' (examples 21 and 23), *pähe jääma* 'retain, remain in memory' (example 22).

(19) *ˋpandi sur kübar **pähe*** (Muh)
'a big hat was **put on**' (lit. 'put onto the head')
(20) *veel temä viruttand ˈroikkaga **pähä*** (Juu)
'and he had whacked with a pole **on the head**'
(21) *siiski ˈlaulusi minul ˈakkas **pähe** ˈkuigi ˈkeegi ˈlaulis* (Vän)
'I could still **memorise** the songs, how somebody sang'
(22) *tean mull omm nüüt **pähä** jäänu* (Kam)
'I know that it has remained in my **memory** now'
(23) *ei no vesi **pähä** ei ˈakka* (Mih)
'water does not go **to** your **head**'

**3.2.4.** TIME. The most frequent words denoting time are *aeg* 'time', *aasta* 'year, *päev* 'day', *kord* 'turn, time', *õhtu* 'evening', and *hommik* 'morning'. Along with object cases, the adessive is also used in time expressions. Two groups can be formed with time words: one group denotes (countable) time units (e.g. *aasta, kord*), which are

generally used only in word combinations; words belonging to the other group can denote time units as well as characteristics of time (*päev, õhtu, hommik*).

The word *aasta* 'year' is used more often in an object case, in the partitive, since it typically modifies quantifiers (example 24). In the genitive the word *aasta* appears with an adposition or as a time adverbial (equivalent in meaning with the adessive, example 25). As a time adverbial, the genitive can also denote duration (example 26).

(24) *suur sõda ol'l nel'i* ****aastat*** (Kam)
'the big war lasted four **years**'
(25) *teese ˈ**voasta** sai teese ˈ**voasta** sai kasukka palgaks* (Muh)
'in the second **year** in the second **year** they paid me with a fur coat'
(26) *pedas ˈvastu kaa möne ˈ**aasta*** (Khk)
'he held on for a few **years**'

The word *kord* 'turn, time' is most often used in object cases (equally frequently in nominative and partitive, the third most frequent form is the genitive). *Kord* may have different meanings ('repeated time moment', 'layer', 'floor', 'order'), but in dialect texts it is mainly used to denote time (examples 27 and 28).

(27) *tuu ol'l' s mittu ˈ**kõrda** Suri mann k'äünüq* (Har)
'he/she had been to Suri several **times**'
(28) *teine **kord** e ˈpüöra teist kätt* (Kei)
'next **time** go to the other direction'

The words *päev, hommik, õhtu* form a separate group among the frequent words referring to time. These words can be used to denote the division of time, but they can also denote a special quality of time and be used as separate phrases. Along with object cases, the adessive form of the words *hommik* and *õhtu* has also made it to the frequency table (the proportion of the adessive forms is con-

siderably smaller compared to the nominative or genitive; in the literary language, however, the adessive is the predominant form).

The proportion of the partitive forms of the word *päev* 'day' can be explained with their use as modifiers of quantifiers (same as *aasta* 'year' and *kord* 'turn, time'). There are some partitive forms among the words labelled as nouns which can be considered as adverbs (example 29). In 20 cases the time adverbial in a partitive form has been labelled as an adverb (examples 30 and 31). This seems justified since a time adverbial with an attribute is usually in the genitive (example 32). A similar two-way labelling was noticed in case of the words *hommik* 'morning' and *õhtu* 'evening'. The nominative singular is also generally used as a time adverbial (example 33).

(29) *nüit e ˈ**päivä** siss ä olli sääl* (Rõn)
'I was there during the **day**'
(30) *lapsed ˈollid ˈikka ˈ**pääva** kodus* (Muh)
'the kids were at home during the **day**'
(31) *siss ˈüüse ˈvahtsõ ˈaida ˈ**päivä** ˈteije tüüd* (Amb)
'then during the night I looked at the garden and during the **day** I was working'
(32) *siis e teese **päävä** ˈtul'di ˈjälle ˈkokko* (Khn)
'then on the next **day** we came together again'
(33) *ja iga **pääv** *pes'ti last* (Kuu)
'and every **day** they washed the child'
Used in the nominative or genitive, the word *päev* can also mean 'päike'.

The genitive and nominative forms of the word *õhtu* 'evening' are normally used as time adverbials. Actually, it is not always possible to distinguish which form has been used and the people who have tagged the texts have used different labels. Referential uses are clearly nominative (example 34). In case of time adverbials the interpretations are different, e.g. nominative singular (example 35)

and genitive singular (example 36). In addition, in 49 cases the form has been labelled an adverb (example 37).

(34) *üks pühaba ˈ**õhta** ilus ˈ**õhta** õl'i* (Trm)
'one Sunday **night**, it was a beautiful **night**'
(35) *vanast ˈmintti ˈlauba ˈ**õhtu*** (Krk)
'in old times people went on Saturday **night**'
(36) *pan'di jõulu ˈlauba ˈ**õhta** keriselle ja ˈahju ˈküpsma* (Pil)
'on Christmas **Eve** they were put on the sauna rocks and into the stove to cook'
(37) *siis nohh siis pühaba ˈ**õhta** siis ehittat'ti noorik ära* (Avi)
'then on Sunday **night** the young wife was decorated'

The nominative/genitive form is predominant in time expressions, the adessive forms characteristic of literary language are relatively rare in the dialect corpus.

Genitive form is the most frequent form of the noun *hommik* 'morning' (example 38). In addition, in 65 cases the time adverbial in the genitive form has been labelled as an adverb (example 39). Adessive form occurs 17 times.

(38) *ja ˈpan'tti siss **hommukku** tul'i ˈahju* (Kam)
'and the fire was lit in the **morning**'
(39) *ja **ommikku** sai nii vara ˈmetsa* (VJg)
'and in the **morning** we went into the woods so early'

In case of the time words like *päev, õhtu, hommik* it can be seen that they have adverbialized. In case of *päev* 'day', the reason to consider the partitive form as an adverb is that a time adverbial with an adjunct is generally in the genitive form, and without an adjunct in the partitive. It is possible that we are dealing with an adverbialized case form which has at one point become detached from the paradigm and which has the same form as the partitive (Velsker 2006: 185). Among the time words, the adverb *ööse* 'at night' has

acquired a shape separate from the noun forms. Since time expressions can be considered as a separate system, the separation of one form from the paradigm allows the other forms to be interpreted as adverbs more easily.

The word *aeg* 'time' is different from the rest of the time words because the most frequent form is the adessive. The frequent use of the adessive forms may indicate the tendency for a word form to become an adposition; the abundance of certain clusters is also predominant – most frequent is the word clusters *sel/tol ajal* 'at this/that time', a construction with a more abstract meaning and which tends to become fixed and could change into an independent (pro)adverb (example 40).

(40) *ja siis siis sell **aal** siss akkat'ti neid nimesid* (Avi)
'and at this **time** these names started to be'

On the other hand, all sorts of different attributes may be added to the word *aeg* (examples 41 to 43). Such cases are limited to adpositions and in 40 cases the word has been categorised as a postposition (with the keyword *ajal* 'during, at'). Thus, the boundary between nouns and postpositions is not clear at all.

(41) *ˈjälle suisel **aal** näd sii eläväd ja* (Var)
'at summer **time** they live here'
(42) **eina+ma jagusi oli minugi **aeal** viel viis jagu oli* (Jõh)
'there were hayfield sections even at my **time**, five sections there were'
(43) *ja ja nüid koloosi **aal** ˈtetti si karjalaut ˈkorda* (Trv)
'and now at the **time** of the collective farms the cowshed was fixed'

The nominative form can be used both referentially (example 44) as well as adverbially (example 45), the last one is similar in its use and meaning to the adessive form *ajal*.

(44) *sügise ol'i kardule ˈvõtmese **aeg** ol'i ˈmihklebäeva laat siis* (Pil)
'at autumn during the **time** of harvesting potatoes it was the Michaelmas fair then'
(45) *aga sõa **aeg** ˈpan'tti põlema* (Äks)
'but **at the time of** the war it was set on fire'

In case of the partitive, the picture is more varied; this form occurs together with expression verbs (*aega võtma* 'take time', *aega teenima* 'serve time', *aega saama* 'find time', *aega viitma* 'spend time') and quantity words (example 46).

(46) *ja tükk ˈ**aega** õl'ime sial* (Trm)
'and we were there for a long **time'**

We can thus say that some time words (*hommik, õhtu, päev*) tend to adverbialize, but the word *aeg* with a more general meaning tends to grammaticalize (to become an adposition).

**3.2.5.** PLACE. Characteristic of words denoting place is the use of internal locative cases – almost all of these words have among the three most frequent forms the illative case (example 47) or the inessive case (example 48). Such words are, for example, *mõis* 'manor', *talu* 'farm', *küla* 'village', *meri* 'sea', *koht* 'place', *ots* 'tip, end', *kari* 'cattle', and also words like *paat* 'boat' and *ahi* 'stove' which denote objects, act as adverbials of space in these forms.

(47) *vanaisa läks oma perega ˈ**metsa** Säre+vere ˈvalda* (Vän)
'grandfather went with his family into the **forest** in Särevere parish'
(48) *noh näq **mõtsahh** el'l'i ja* (Vas)
'well, they lived in the **forest** and'

It can be seen in case of the word *küla* 'village' that certain forms are becoming separate from the rest of the paradigm. The illative form of *küla* has become fixed in the expressions *külla minema* '(go) visit' (lit. 'go into the village') and *külla tulema* '(come) visit' (lit. 'come into the village'). Some of these cases have been labelled as affixal adverbs (being part of expression verbs, example 49), but most as nouns. Nevertheless, it is clear that this kind of use no longer refers to the meaning of the noun 'village', but has diverged from it and become a fixed part of the expression verb, i.e. lexicalized.

(49) *läts' tõsõlõ talolõ ˈ**küllä*** (PSe)
'he went to visit another farm'

The words *maa* 'land, ground' and *tee* 'road, path' differ from the other words in their use of cases. For *maa* the most frequent are the genitive and partitive forms. The frequent use of the genitive can be explained by its occurrence together with adpositions (example 50) and its use as an attribute. The partitive, again, is often used to characterise certain quantitative parameters, including its use as a modifier in quantity phrases (example 51). In the nominative, the word *maa* in its different meanings is used as a subject or an object.

(50) *vahest sai kohe **maa** *külge *kinni *pandud* (Hlj)
'maybe we managed to fix it to the **ground** at once'
(51) *ˈraasukke **maad** ˈkaugemal* (Äks)
'a bit further away' (lit. 'for a bit of **land** further')

The frequent use of the genitive forms of the word *tee* refers to their use together with adpositions (example 52).

(52) *siss üt's jäi sinnäq **tii** pääle sääl om rist* (Har)
'then one of them remained on the **road**, there is a cross'

In addition to the words denoting concrete places (*mõis, talu, mets, küla, maja*), other frequently used words may acquire the meaning of place. Probably due to the specific texts in the dialect corpus, the inessive forms of the word *kari* 'cattle' which denote rather a place (and also the activity, example 53) are among the most frequent noun forms.

(53) *isa käis* ***karjas*** (Lüg)'
'father tended cattle' (lit. 'father went in the **cattle**')

Among the forms of the word *koht* 'place', the inessive is also the most frequent one. This word may have either an abstract or a concrete meaning ('farm place', example 54), the general meaning dominates with the inessive forms and they usually include more abstract place references (example 55). With nominative forms one can also find concrete meaning, e.g. *talukoht* 'farm place' (example 54).

(54) *poissmiis ja siss jäi* ***koht*** *selle teise venna 'kätte* (Äks)
'bachelor and then the **place** was left for another brother'
(55) *mõnes* ***kohas*** *küll *teises* ***kõhas*** *oli kohe niij ett*... (Jõh)
'in some **places** it was so, but in other **places** it was so that...'

It can be seen when analysing the material in the dialect corpus that for certain word forms the decision of whether it is a noun or an adverb or an adposition depends on the person who has tagged the text. On the one hand it could be said that the principles of tagging need specification, but on the other hand it denotes language change, transition areas and the fuzzy boundaries between word classes. We can take the word *ots* 'tip, end' as an example: 84 (20.3%) among the forms labelled as nouns are inessive forms, at the same time, the same word form has been labelled as an adverb in 42 cases, as an affixal adverb in 8 cases and as a postposition in 71 cases. The big proportion of the inessive forms can be explained by grammaticalization (these forms are becoming adpositions) –

a word form has been labelled a postposition when the concrete meaning has become vague and semantic bleaching is one of the characteristics of grammaticalization (Heine 2005) (example 55); but even the same word clusters have been labelled differently (also when the concrete meaning still exists; compare examples 56 where there is a noun and example 57 where there is a postposition).

(55) *piim ka l'l käe* ***otsah*** (Vas)
'he also had milk in his hand' (lit. 'in/on the tip of his hand')
(56) *sie õli *pikka *varre* ****õtsas*** (Lüg)
'it was on a long stalk' (lit 'in/on the tip of the stalk')
(57) *tuu l'i sääratse varre* ***otsan*** (Nõo)
'that was on a such a stalk' (lit 'in/on the tip of the stalk')

Overlap with nouns also occurs in labelling the form as an adverb – most clear cases of adverbs are those where there has been a meaning transfer (*otsas* meaning 'through, gone', example 58).

(58) *teene leib oli jo* ***otsas*** *eij old ˈsöija* (Amb)
'the other bread was all **gone**, there was nothing to eat'

## 4. Conclusion

In this article we gave an overview of the dialect corpus and its state in August 2008. We furthermore analysed the most frequently used nouns and the use of their most frequent forms. We found that the list of the most frequent words in the dialect corpus is similar in some ways to the frequent vocabulary of the modern literary corpus. Nevertheless, the dialect corpus includes words which derive from the specificity of the texts – these texts relate to the traditional rural culture.

It was seen from the analysis of the frequent word forms that the frequent use of some word forms can easily be explained by their use as core arguments of a clause – as a subject or an object. Such words were those denoting people and other living beings (*mees, naine, inimene, laps,* etc.) in which case the most frequent forms were the nominative, genitive and partitive. In case of many words the accumulation of certain forms may be noticed; for example for words denoting time the genitive or partitive is often used as a time adverbial. In such cases the form is no longer as transparent as it used to be and it shows signs of adverbialization. The people who have tagged the corpus have also used the adverb tag more often in such cases instead of the noun tag. In case of time words, the adessive was also frequently used.

The body part terms tend to grammaticalize (e.g. *käsi* 'hand') and in the texts we could find instances of grammaticalization to lesser or greater extent. The body part terms had also lexicalized in certain forms, become fixed as certain adverb forms, which are often related to clothing (*pähe* 'head' sg ill).

In locative expressions, the interior locative cases were often used (inessive and illative). Thus, it can be said that the exterior locative cases express more abstract relations (time, possession), but the interior locative cases more concrete relations (place).

Certain word forms show signs of grammaticalization, i.e. signs of becoming adpositions (e.g. *otsas* inessive singular of *ots* 'end, tip', *ajal* adessive singular of *aeg* 'time'), but the extent of grammaticalization is not always clear. The texts include lesser or more grammaticalized forms, where the people who have tagged the texts have found it difficult to judge whether to label the word as a noun form or some other already grammaticalized word class. The same sort of confusion exists on the borderline between adverbs and nouns – certain very frequent forms of certain words become autonomous and it is no longer clear whether they belong to the noun paradigm or not. Such were especially time expressions (*hommiku* singular genitive of *hommik* 'morning', *õhtu* singular genitive

and nominative of *õhtu* 'evening'); the people who have tagged the texts have also had hard time with these words.

It is important from the point of view of the dialect corpus to specify the boundaries between different word classes; on the other hand, this is not always possible, because sometimes the boundaries are fuzzy. This is primarily due to the processes of lexicalization and grammaticalization, which have not come all the way yet. Language, as well as dialect language, is constantly changing.

## Abbreviations

| | |
|---|---|
| 3 – 3rd person | ipf – imperfective |
| ad – adessive | kom – comitative |
| el – elative | n – nominative |
| g – genitive | p – partitive |
| ill – illative | pl – plural |
| in – inessive | ps – personal |
| ind – indicative | sg – singular |

**Estonian parish dialects**

Amb–Ambla (Mid); Avi–Avinurme (Eastern); Har–Hargla (Võru); Hlj–Haljala (Coastal); Hää–Häädemeeste (Western); Jõe–Jõelähtme (Coastal); Jõh–Jõhvi (North-Eastern); Juu–Juuru (Mid); Kam–Kambja (Tartu); Kei–Keila (Mid); Khk–Kihelkonna (Insular); Khn–Kihnu (Insular); Krk–Karksi (Mulgi); Kuu–Kuusalu (Coastal); Lüg–Lüganuse (North-Eastern); Mih–Mihkli (Western); Muh–Muhu (Insular); Mus–Mustjala (Insular); Nõo–Nõo (Tartu); Pil–Pilistvere (Mid); PSe–Põhja-Setu (Seto); Rõn–Rõngu (Tartu); Trm–Torma (Eastern); Trv–Tarvastu (Mulgi); Var–Varbla (Western); Vas–Vastseliina (Võru); VJg–Viru-Jaagupi (Mid); Vän–Vändra (Western); Äks–Äksi (Mid)

## References

Bybee, Joan 2005. Mechanisms of Change in Grammaticalization: The Role of Frequency. – Joseph, B.D.; Janda, R.D. (eds.) The Handbook of Historical Linguistics. Oxford: Blackwell Publishing, 602–623.

Bybee, Joan 2007. Frequency of Use and the Organization of Language. Oxford: Oxford University Press.

EKG I = Erelt, Mati; Kasik, Reet; Metslang, Helle; Rajandi, Henno; Ross, Kristiina; Saari, Henn; Tael, Kaja; Vare, Silvi. Eesti keele grammatika I. Morfoloogia. Sõnamoodustus. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut 1995.

Erelt, Mati 2003. Syntax. – Erelt, M. (ed.) Estonian Language. (Linguistica Uralica Supplementary Series Vol. 1. Tallinn: Estonian Academy Publishers, 93–129.

Heine, Bernd 2005. Grammaticalization. – B.D. Joseph, R.D. Janda (eds.) The Handbook of Historical Linguistics. Oxford: Blackwell Publishing, 575–601.

Hennoste, Tiit 2002. Suulise kõne uurimine ja sõnaliigi probleemid. – Pajusalu, R.; Tragel, I.; Hennoste, T.; Õim, H. (eds.), Teoreetiline keeleteadus Eestis. (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4.) Tartu: Tartu: Ülikooli Kirjastus, 56–73.

Hennoste, Tiit 2004. Mõnede käänete sagedus ja lauseliikmelisus suulises kõnes. –Lindström, Liina (toim.), Lauseliikmeist eesti keeles. (Tartu Ülikooli eesti keele õppetooli preprindid 1.) Tartu: Tartu Ülikooli Kirjastus, 16–25.

Hopper, Paul J., Elizabeth C. Traugott 2003. Grammaticalization. Second Edition. Cambridge: University Press.

Kaalep, Heiki-Jaan, Kadri Muischnek 2002. Eesti kirjakeele sagedussõnastik. Tartu: Tartu Ülikooli Kirjastus.

Kasik, Reet (toim.) Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks. (TÜ eesti keele õppetooli toimetised 17.) Tartu: Tartu Ülikooli Kirjastus, 212–221.

Lindström, Liina 2001. Eesti murrete korpuse iseloomustus argivestlustega võrrelduna. –

Lindström, Liina 2005. Ülevaade Eesti murrete korpusest. www.murre.ut.ee/EMK.PDF

Lindström *et al.* 2001 = Lindström, Liina; Lonn, Varje; Mets, Mari; Pajusalu, Karl; Teras, Pire; Veismann, Ann; Velsker, Eva; Viikberg, Jüri 2001. Eesti murrete korpus ja kolme murde sagedasema sõnavara võrdlus. – Kasik, Reet (toim.) Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks. (TÜ eesti keele õppetooli toimetised 17.) Tartu: Tartu Ülikooli Kirjastus, 186–211.

Lindström *et al.* 2006 = Lindström, Liina; Bakhoff, Liisi; Kalvik, Mari-Liis; Klaus, Anneliis; Läänemets, Rutt; Mets, Mari; Niit, Ellen; Pajusalu, Karl; Teras, Pire; Uiboaed, Kristel; Veismann, Ann; Velsker, Eva 2006. Sõnaliigituse küsimusi eesti murrete korpuse põhjal. – Niit, Ellen (toim.) Keele ehe. (Tartu Ülikooli eesti keele õppetooli toimetised 30.) Tartu: Tartu Ülikooli Kirjastus, 154–167.

Pajusalu, Karl 2003. Estonian Dialects. – M. Erelt (ed.) Estonian Language. Linguistica Uralica Supplementary Series Vol. 1. Tallinn: Estonian Academy Publishers, 231–272.

Pajusalu *et al.* 2002 = Pajusalu, Karl; Hennoste, Tiit; Päll, Peeter; Viikberg, Jüri 2002. Eesti murded ja kohanimed. Tallinn: Eesti Keele Sihtasutus.

Valge, Jüri 1970. Eesti keele käänete sagedused kolmes funktsionaalses stiilis.– Keel ja Struktuur 4. Tartu: Tartu Riiklik Ülikool, 145–162.

Velsker, Eva 2006. Ööd tähistavad ajaväljendid eesti murretes, ühis- ja kirjakeeles. – Emakeele Seltsi aastaraamat 51. Tallinn: Eesti Teaduste Akadeemia Emakeele Selts, 184–207.

# *Mees*, *aeg* ja teised sagedased sõnad Eesti murrete korpuses

Liina Lindström, Eva Velsker, Ellen Niit, Karl Pajusalu

## Resümee

Artiklis tutvustatakse Tartu Ülikooli ja Eesti Keele Instituudi koostöös koostatavat Eesti Murrete Korpust ning selle hetkeseisu ning analüüsitakse korpuses esinevat sagedasemat sõnavara. Vaatluse all on 100 kõige sagedasemat substantiivi ning nende substantiivide kõige sagedasemad vormid, mis on saadud murdekorpuse morfoloogiliselt märgendatud tekstidest.

Murdekorpuses esinevate substantiivide sagedusloendit võrreldakse tänapäeva kirjakeele korpuse samataoliste andmetega. Selgub, et suur osa sõnavara on murdekorpuses sama, mis kirjakeele korpuses, hoolimata sellest, et esimene esindab 1960ndatel lindistatud ja valdavalt 19. sajandil sündinud inimeste keelt, viimane aga 20. sajandi lõpu kirjakeelt. Suured erinevused tulenevad muutunud sotsio-kultuurilistest asjaoludest: murdekorpuse tekstid kajastavad 19. sajandi lõpu – 20. sajandi alguse eesti agraarkultuuri (nt sõnad *mõis, talu, küla; hobune, lehm, piim* jne).

Osa murdekorpuse sõnavara on sagedusloendites aga raskesti selgitatavad sotsiokultuuriliste olude vms abil (nt *ots*). Eelkõige just nende vormide selgitamiseks oleme analüüsinud 50 sagedasema substantiivi kolme kõige sagedasemat vormi, et näha, kas toimub teatud vormide kuhjumine, mis laiemalt on seotav leksikaliseerumis- või grammatiseerumisprotsessiga.

Sagedasemate sõnavormide analüüsist selgub, et osa sagedasemaid sõnavorme on põhjendatavad nende kasutamisega lause tuumargumentidena – subjekti ja objektina. Sellised olid isikutele ning muudele elusatele ja elututele objektidele viitavad sõnad (*mees, naine, inimene*, laps jne), mille puhul kasutati enim nomina-

tiivi, genitiivi ja partitiivi. Paljude sõnade puhul võib aga märgata teatud vormide kuhjumist – näiteks aega tähistavate puhul kasutatakse sageli genitiivi (*hommiku, õhta*) või partitiivi (`*päeva*) ajamäärusena, nende puhul ei ole vorm enam nii läbinähtav ning võime märgata adverbistumist.

Teatud sõnavormide puhul (nt *otsas*, *ajal*) on märgata vormi grammatisatsiooni kaassõnaks, ent alati ei ole selge, mil määral grammatisatsioon on toimunud – tekstides on nii rohkem kui vähem grammatiseerunud vorme, mille puhul märgendajatel on raske hinnata, kas märgendada sõna substantiivivormiks või juba kaassõnaks. Sama segadus on ka adverbi ja substantiivi piirimail – teatud sõnade teatud väga sagedased vormid muutuvad autonoomseteks ning nende kuulumine substantiiviparadimagasse ei ole enam ilmne. Sellised olid eriti ajaväljendid (*hommiku* sg g, *õhtu* sg g, sg n), ka märgendajatel on nendega seoses olnud palju segadust. Seda ei saa aga pidada otseselt märgendajate või märgendussüteemi vigadeks, vaid see on tingitud keeles pidevalt toimuvatest muutustest.