Usage patterns of Estonian experience perception verbs: A random forest approach

Mariann Proos
University of Tartu

Abstract

The present study looks at the usage patterns of five Estonian perception verbs (*nägema* 'to see', *kuulma* 'to hear', *tundma* 'to feel', *maitset tundma* 'to taste' and *lõhna tundma* 'to smell') in a corpus sample. The aim is to describe the similarities and differences that emerge in their usage patterns, as well as determine which contextual cues could be most useful for determining the meaning of the verb in the sentence. A corpus sample of 2,418 sentences was annotated and a random forest model was constructed to reach these aims. The random forest model shows that construction type is the most informative variable in predicting the verb in the sentence, followed by semantic type of stimulus and morphological person. Each of the most significant variables is discussed separately in terms of how they inform the meaning of the verb in a sentence. The study concludes that in Estonian, a language with rich morphosyntax, form and meaning are innately linked, with no causal link between the two.

Keywords: Estonian, random forest, corpus, perception verbs

1 Introduction

Languages use different strategies to lexicalise various kinds of perception experiences that are related to our five senses: sight, hearing, touch, gustation and olfaction. In his seminal work, Viberg (1984) laid out three types of what he calls *dynamic modalities of perception verbs*: experience perception verbs (e.g. 'to see'), activity perception verbs (e.g. 'to look'), and source-based perception verbs (e.g. 'to look'; *you look nice tonight*). The paradigm of perception verbs thus contains five modalities and three dynamic types, offering a total of 15 possible different lexical elements to cover the realm of main perception verbs.[1] Some languages exhibit total distinction, where each

[1] Main perception verbs are verbs that are semantically neutral, i.e. they have no connotation and are not marked for lexical aspect.

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

"slot" is filled by a different lexical item. Other languages have total lexical differentiation in some of the sense modalities (typically seeing and hearing), but use one verb across the other modalities and dynamic types (Viberg 1984).

Languages can also exploit different morphosyntactic strategies to lexicalise the different types of perception. For example, some languages like Swedish, Finnish, and Estonian (i.a.) use compound verbs to lexicalise the experiences of smelling and tasting (Viberg 1984: 143). In these languages, the experience verbs for smelling and tasting are comprised of the tactile perception verb and a noun phrase that denotes smell or taste respectively (e.g. *lõhna/maitse-t tundma* smell.PART/taste-PART feel.INF2 'to smell/taste'). The present study looks at the usage patterns of five Estonian experience perception verbs: *nägema* 'to see', *kuulma* 'to hear', *tundma* 'to feel', *maitset tundma* 'to taste', and *lõhna tundma* 'to smell'.

On the level of argument realization, Croft (2012) has grouped emotion, cognition, and perception as mental events and has showcased how the argument structure patterns are similar across languages. He posits two main realization patterns: 1) the experiencer is realized as the Subject, and the stimulus is realized as the Object or Subsequent Oblique;[2] 2) the experiencer is realized as the Object or Subsequent Oblique, and the stimulus is realized as the Subject (Croft 2012: 233). Carrying this over to Viberg's dynamic system, this is the distinction between the *experience dynamic system* and the *source-based dynamic system* (Viberg 1984: 128). For example, the sentence *Mary saw John* is an expression of the experience dynamic system, where *Mary* is the experiencer and is realized as the subject, and *John* is the stimulus and is realized as the object. As the present study looks at experience perception verbs, the stimulus is the participant in the sentence that would exhibit grammatical variation according to Croft (2012). Thus, the present study pays special attention to the encoding of the stimulus.

As all of the verbs included in the study belong to the same class, experience perception verbs, it is reasonable to assume that they should exhibit behaviour that is similar to some degree in a corpus sample. They should bear some similarities in what types of participants they code in an event construal, and how these participants are lexicalised. For example, for all of the verbs, the experiencer should be a mandatory participant; the stimulus can but need not be lexicalised based on Croft's (2012: 233) overview. On the other hand,

[2] A *Subsequent Oblique* denotes an event that is subsequent to the one expressed by the verb (Croft 2012: 276–277).

semantically, they lexicalise different experiences, and also exhibit different degrees of polysemy. Therefore, individual differences between the verbs should also be observable from a corpus sample. For example, while *tundma* 'to feel' has been observed to lexicalise a variety of emotional and mental experiences (Proos 2020b), *nägema* 'to see' has a much more limited semantic extension potential (Proos 2019). Thus, we are dealing with a collection of verbs that a) should have similar meanings and similar usage patterns because they belong to the same verb class, but b) should exhibit some differences in meaning and differences in usage patterns, since they lexicalise different types of experiences and have varying degrees of polysemy.

This study focusses on the similarities and differences of the variation of usage patterns that Estonian experience perception verbs exhibit. The notion of *usage pattern* here refers to potential variation on both the semantic and the morphosyntactic level. At the core of the present study is the distributional hypothesis (Harris 1954). According to the hypothesis, the meaning of a linguistic unit is dependent on the context it occurs in. Words that occur in similar contexts thus have similar meaning – words that appear in different contexts have meanings that are less similar. For example, the words *cat* and *dog* are likely to occur in more similar contexts than *bus* and *dinosaur*. Derived from this, given correct and sufficient contextual cues, the language element should be predictable from context alone. This is not to say that the language user makes their choices purely on the basis of contextual cues, by inserting an appropriate verb into the "verb slot". However, there is a degree of frequency and co-occurrence information that goes into the process of language production and comprehension.

A strong tradition of corpus-based research that combines both semantic and morphosyntactic information has been established in recent years that takes the distributional hypothesis as its basis. Numerous studies have looked into alternation on the syntactic level (Bresnan et al. 2007; Bresnan & Ford 2010; Grafmiller & Szmrecsanyi 2018; Szmrecsanyi et al. 2019; Klavan 2020) as well as the lexical level (Peirsman et al. 2010; Franco et al. 2019). One of the methods, the behavioural profile method, has also proven to be useful for meaning research for both near-synonymy (Divjak 2010) and polysemy (Gries 2006; Berez & Gries 2008; Glynn 2014; 2016). Using what they call *constructional profiles*, Janda & Solovyev (2009) found that it is possible to distinguish between near-synonymous nouns for HAPPINESS and SADNESS in Russian. Divjak (2015) found that it is possible to differentiate between perception verbs of seeing, hearing and touching in Russian based

on only morphosyntactic variables. Thus, the distributional approach has been shown to be a suitable basis for describing language on its various levels, from morposyntax to semantics. The present paper largely follows the methodology adapted by Divjak (2015) in her study concerning the grammatical variation of Russian perception verbs.

The bulk of the work until now has been done on Indo-European languages that exhibit less morphosyntactic variation than languages from some other language families (e.g. Finno-Ugric languages). Finno-Ugric languages offer a unique window into the numerous possibilities of morphosyntactic variation and its relation to semantics. The present study aims to expand the knowledge we have about Finno-Ugric languages in this respect by using data from Estonian.

To sum up, the present study looks at Estonian experience perception verbs and their distribution in a corpus sample. The aims of the study are to:

1. explore how verbs of the same class (experience perception verbs) but of different morphosyntax (simple vs. compound verbs) behave in a corpus sample;

2. determine which contextual elements would best allow the speaker to derive the meaning of the perception verb in the sentence.

Perception verbs in general tend to be polysemous in languages. They are often used to lexicalise non-perceptual, abstract experiences in addition to sensory experiences. There are some meaning patterns that are universal across languages, e.g. a general PERCEPTION → UNDERSTANDING metaphor is postulated by Ibarretxe-Antuñano (2008). However, different languages use the vocabulary of different sensory fields as a source for the metaphor. For example, Australian languages have been shown to use audition vocabulary as the source (Evans & Wilkins 2000), English and other Indo-European languages use seeing vocabulary as the principle source for this extension (Sweetser 1990), and in Finnish, Swedish (Viberg 2015) and Estonian (Proos 2020b), tactile perception vocabulary is the primary source for the extension.

The explanatory dictionary of Estonian, *Eesti keele seletav sõnaraamat* (EKSS) (Langemets et al. 2009), lists 8 meanings for *nägema*, 5 meanings for *kuulma*, and 11 meanings for *tundma*. Furthermore, Proos (2019) has analysed 13 different meanings of *nägema* 'to see' via a sorting task and a behavioural profile analysis, and Proos (2020b) offers an experimental approach to the polysemy of *tundma* 'to feel' which includes 25 different polysemous senses

of the verb. Polysemy certainly plays a large role in the usage patterns of the verbs under study. Looking at the usage patterns of the verbs also informs us about the polysemy of the verb and the role that polysemy plays in variation of the verbs. Since meaning is a function that can be assessed with the means of a distributional approach to language, we should see at least some of the different meanings emerge in the usage patterns of the verbs; e.g. a specific type of perception stimulus could indicate a specific type of meaning of the verb in question.

If, as outlined in the previous paragraphs, context can be used to predict the linguistic element in a sentence out of variants, it should be possible to create a model that predicts the choice of perception verb. However, it is not clear which contextual information is the most relevant and sufficient for this purpose. To answer the research questions, a corpus sample was annotated for various semantic and morphosyntactic information. A random forest model was then compiled to find out which of the predictors are the most significant in determining the choice of the verb in the sentence. In the following sections, the material and method will be more closely introduced (§ 2). This is followed by § 3, where the results of the modelling are discussed. A general discussion follows in § 4.

2 Material and methods

The five Estonian verbs included in the study are as follows: *nägema* 'to see', *kuulma* 'to hear', *tundma* 'to feel', *maitset tundma* 'to taste' and *lõhna tundma* 'to smell'. The latter two verbs are compound verbs, composed of the tactile perception verb *tundma* 'to feel' and the nouns *maitse* 'taste' and *lõhn* 'smell' respectively.[3] The compound verbs are separable, i.e. the verb and the noun can appear in different parts of a sentence – they need not be in adjacent positions.

There are two possibilities of analysing the *maitset tundma* 'to taste' and *lõhna tundma* 'to smell' expressions: on the one hand, these are a simple verb + direct object constructions, where the verb *tundma* 'to feel' takes a direct object in the partitive case (which, together with the genitive, is a typical grammatical object case in Estonian). On the semantic level, this analysis

[3] The nouns *maitse* 'taste' and *lõhn* 'smell' can occur in the partitive or nominative case. The nouns take the nominative case only when the compound verb is used in a specific passive construction. This is discussed in § 3.1.

would suggest that combining with the nouns *maitse* 'taste' and *lõhn* 'smell' is just a possible characteristic of the verb *tundma* 'to feel', and 'to taste' and 'to smell' are polysemous meanings of *tundma* 'to feel'.

On the other hand, the expressions *maitset tundma* 'to taste' and *lõhna tundma* 'to smell' are lexical units that express the perceptual experiences of tasting and smelling. If we consider these in the framework of the main perception verb paradigm (Viberg 1984: 125), they can be analysed as dedicated lexical units for expressing a certain type of perception, and thus not simply as polysemous meanings of *tundma* 'to feel'. I believe these are not mutually exclusive approaches. Since the goal of this paper is to look at the experience perception verbs in Estonian, the second approach is mainly preferred, i.e. *maitset tundma* 'to taste' and *lõhna tundma* 'to smell' are considered complete lexical units and named *compound verbs* here. However, their grammatical composition is still considered as a case of a verb combining with a direct object. For example, this is reflected by the variable "construction type" in the present study, where the syntactic composition of the perception verb constructions in the corpus sample is considered.

For the analysis, 500 sentences per perception verb were collected from the Estonian Web Corpus (etTenTen 2013),[4] with the exception of *maitset tundma* 'to taste', where 418 sentences remained after manually checking the sample. The full sample included 2,418 sentences.[5]

2.1 Variables

Within the corpus-based meaning research tradition, there is an established tradition of what types of variables can be included in the annotation schema. However, the final choice of variable types is dependent on the specifics of the sample at hand as well as the research questions. The annotation of variables is discussed at length for example in Divjak & Gries (2006), Divjak & Fieller (2014), Gries (2006) and Glynn (2014; 2016). Traditionally, the variables include both morphosyntactic and semantic variables. In the present study, the morphosyntactic variables included follow the example of Divjak (2015). In addition, semantic variables are included since the goal of the present study

[4] The corpus consists of 686,000 web pages in Estonian with a total of 270 million words. This corpus represents a very wide range of text types, ranging from forum texts and newspapers to scientific texts.

[5] The material and analysis script along with extra materials for this study is accessible through Open Science Framework (Proos 2020a).

is to look at semantic variation as well.

The present paper takes a bottom-up approach to variable annotation. This means that the variable levels are not decided upon *a priori*, but arise from going through the data. This allows for a flexible presentation of the full range of semantic and morphosyntactic variation that the verbs exhibit in the corpus sample. However, this also means that variation that goes beyond what is traditionally considered under the word class *verb* is also included in the study. For example, instances of perception verbs being used as particles are also included in the study, although traditional Estonian grammar does not consider these forms as verbs (Erelt & Metslang 2017: 58).

All of the sentences were automatically tagged for morphological information (person, tense, mode, polarity, voice) in regards to the perception verb in the sentence. Some sentences were not tagged automatically, since there is considerable grammatical homonymy in Estonian: these sentences were checked manually. For example, Estonian uses the connegative verb form in negation so person is not marked on the verb (*ma ei näe* 'I don't see' vs *te ei näe* 'you (2PL) don't see'). In addition, two semantic and one syntactic variable were manually coded.

As can be seen from Table 1, the sense of the verb was not included as one of the variables. This was done for the following reasons. Firstly, polysemy is considered here as variation of the verb's meaning and thus a part of the different usage patterns a verb potentially exhibits. Using the methodology outlined in this section, we should see some patterns emerge that represent this type of meaning variation. Secondly, on a methodological level, annotating senses in a corpus sample is problematic due to the nature of polysemy itself. As Glynn (2016) points out, assigning a sense for each sentence would presume that polysemous senses are discrete units. However, according to the theory of cognitive semantics, senses should be treated as categorical in nature, with areas of overlap. Lastly, Glynn (2016) also points out that since the researcher needs to rely on contextual information to be able to assign a sense to the verb, this results in a disproportionately large correlation of variables, i.e. the researcher is depending on other variables to annotate the sense variable.

As Table 1 shows, the two semantic variables concern the stimulus of the perception. In the present paper, the term *stimulus* is used to denote the source of perception, i.e. the entity that is being perceived. This can be encoded differently on the syntactic level for different perception verbs. e.g. compare the sentences in examples (1–3) (stimuli marked in boldface):

Table 1. Variables used in the corpus study

Variable type	Variables	Variable levels
VERB MORPHOLOGY	PERSON	1,2,3 singular; 1,2,3 plural; infinitive; passive past participle; active past participle, impersonal present, impersonal past
	POLARITY	affirmative, negative
	TENSE	past, present
	MODE	indicative, imperative, conditional, quotative
	VOICE	personal, impersonal
CONSTRUCTION TYPE	CONSTRUCTION TYPE	verb in intransitive phrase, verb used as adjectival participle, verb + object, verb + adverbial, verb + object + adverbial, verb + object + adverb, verb + clause, verb + adjectival modifier + object, verb + nominal modifier + object, verb + nominal modifier + adjectival modifier + object, verb + reflexive particle + adverb, verb + reflexive particle + noun, verb + particle + object, verb + particle + adverbial
SEMANTICS	SEMANTIC TYPE OF STIMULUS	abstract, ambivalent, concrete, person, sense, situation/event
	ABSTRACTNESS OF STIMULUS	abstract, medium abstract, medium concrete, concrete

(1) *Sookilpkonn peitu-vat mutta iga pisema krabina peale,*
bog.turtle.NOM hide-QUOT mud.ILL each tiny.COMP.GEN rustle.GEN onto
sellepärast on-gi ***teda*** *peaaegu võimatu näha.*
because.of.this be-CLITIC **he.PART** almost impossible see.INF1

'The bog turtle is said to hide in the mud for even a small rustle, that is why it is almost impossible to see **him**.'

(2) *Läksi-n kööki, sest tund-si-n* ***söögi lõhna****.*
go.PST-1SG kitchen.ILL because **feel-PST-1SG food.GEN smell.PART**

'I went to the kitchen because **I smelled food** (lit. feel the smell of food).'

(3) *Kõik me oleme kuul-nud* ***veidra-st lastetõu-st*** *nime-ga*
everyone we be.1PL hear-APP **weird-ELA child.breed-ELA** name-COM
"indigolapsed".
indigo.children

'Everyone has heard about the **weird breed of children** called "indigo children".'

Only in example (1) is the stimulus of the perceptual act encoded as the syntactic object (*teda näha* he-PART see.INF1). In traditional grammar, only objects in genitive and partitive are considered direct objects. In example (2), the syntactic object is the phrase *söögi lõhna* 'smell of food'. However, since the noun *lõhn* 'smell' is a necessary semantic component of the compound verb *lõhna tundma* 'to smell', when analysing the stimulus of the perceptual act, it is more informative to focus the analysis on the noun modifier *söögi* food.GEN, which expresses the source of the scent.[6] Sentence (3) has no syntactic object at all; the elative phrase is an adverbial according to the Estonian reference grammar (Erelt & Metslang 2017). Semantically, however, the elative phrase encodes the stimulus and can thus be considered as an argument of the verb.

2.1.1 Semantic type of stimulus

This variable was included to describe the nature of the stimulus of perception in the sentence. The variable has six levels: abstract, concrete, sense, person,

[6] As was noted by one of the reviewers, it is questionable whether for the verbs *maitset tundma* 'to taste' and *lõhna tundma* 'to smell' the stimulus should be considered as the source of the scent or the scent itself (i.e. is the scent itself what is being perceived or is the thing emitting the scent being perceived). While there is no definitive answer to this, in the present paper, it was decided that for these verbs, the source emitting the scent or taste is the stimulus.

ambivalent, and situation/event. The category "abstract" refers to entities that do not exist physically and are difficult to define (e.g. love, pain, experience, goal, reason). The category "ambivalent" is reserved for stimuli where the meaning is not clear: this mostly includes deictics, pronouns and generic nouns (e.g. that, thing, something, a piece).

Entities that are categorised under "concrete" are entities that can be defined and are mostly also physical, but that cannot be perceived with the sense that the verb in the sentence lexicalises (e.g. world, result, money, work). The entities that can be perceived with the sense that the verb in the sentence lexicalises are categorised with the level "sense". In addition, the categories "person" and "situation/event" are used to categorise persons and events respectively.

2.1.2 Abstractness of stimulus

This variable again concerns the stimulus of the perceptual act. All of the stimuli in the 2,418 sentences were given ratings on a concreteness–abstractness scale and were then split into groups according to their value. Aedmaa (2019) offers abstractness ratings for over 200,000 words in Estonian; the ratings were assigned via semantic vectors (see Aedmaa 2019 for full description of the process). Although care has to be taken because the current ratings have not been assessed by speakers, including the ratings offers an objective value to the description of abstractness and concreteness of the stimuli – something that is notoriously difficult to decide upon. Furthermore, evidence from German shows that ratings based on vector semantics coincide with human ratings to a reliable extent (Köper & Schulte im Walde 2016).

All of the stimuli were manually extracted from the sentences. They were then lemmatised and matched with an abstractness rating based on Aedmaa (2019). There were some stimuli that had no match in the Aedmaa (2019) list: in these cases, the closest possible match was selected, e.g. *motosportlane* → *mootorisportlane* ('moto-athlete' → 'motor athlete'). The Aedmaa (2019) scale offers ratings from 0–10, with 0 being the most abstract and 10 the most concrete. The ratings for the stimuli in the present study were divided up as follows: < 3 abstract; 3–5 medium abstract; 5–7 medium concrete; > 7 concrete. Since some of the stimuli consisted of phrases rather than single words, the average rating of these phrases was considered as the abstractness rating of that stimulus. For example, in the sentence *Franz nägi abilinnapea*

pilti 'Franz saw the picture of the vice mayor' the abstractness rating was assigned to the phrase *abilinnapea pilti* by averaging the ratings for the nouns *abilinnapea* 'vice mayor' and *pilt* 'picture'. The average rating according to the Aedmaa (2019) scale was 7.445, which means the stimulus in this sentence was annotated as "concrete". For sentences where the stimulus of perception was expressed by another clause, no abstractness rating was calculated.

2.1.3 Construction type

Following the example of Divjak (2015), the form of the construction was included as a syntax variable (*argument structure* in Divjak 2015). Since the annotation was conducted with no *a priori* categories in mind, there was no set width of context that was considered as a possible construction. There were many possible construction types and the number of types also varied from verb to verb.[7] The verb phrase was mostly considered as the focal unit, e.g. in a phrase such as *kuulsin sinust juttu* 'I heard a story about you' the construction was annotated as "verb + adverbial + object"; but in a phrase such as *kuulsin sinust juttu eile hommikul oma toredalt naabrilt* 'I heard a story about you yesterday morning from my lovely neighbour' still only the "verb + adverbial + object" was annotated since the other phrases were not considered as crucial components of verb meaning. As the variation in different adverbials is large, the aim was to only include compulsory adverbials in the focal unit; however, as also pointed out by Erelt & Metslang (2017: 300), this distinction is not a straight-forward one. Thus, there remains a degree of subjectivity to the decision of what constitutes the focal unit in the sentence. Altogether 14 different types of constructions across the verbs were annotated, the list of which is provided in Table 1.

The stimulus was marked as "object" when the noun phrase was in the genitive or partitive case. Object-like elements in other cases are considered adverbials, e.g. in the phrase *hooli-n sinu-st* (care-1SG you-ELA) 'I care about you', 'you' is not analysed as an object since it is in the elative case. Thus, *sinu-st* 'you-ELA' was classified as an adverbial.

The past passive participle form in Estonian has an adjective function and thus these tokens are marked as "verb used as adjectival participle". The adjectival participle does not, however, behave like an adjective; for

[7] A number of constructions that were observable from the data are also described in Rätsep (1978) or bear considerbale resemblance to to the clause level constructions that Rätsep (1978) has observed.

example, it exhibits no agreement with noun phrases (*ilusa-te-le mees-te-le* handsome-PL-ALL man-PL-ALL 'for the handsome men' vs. *tuntud mees-te-le* feel.PPP man-PL-ALL 'for the famous men').

Particle verbs were not manually eliminated from the sample. In Estonian, particle verbs are verbal units where the main semantic component is the meaning of the verb. The verb is combined with an adverb, which can have a locative, perfective, state or modality meaning (Erelt & Metslang 2017: 104–106). The ability of verbs to take certain particles and combine into new lexical units is an inherent characteristic of the verb's behaviour and is informative of the nature of the verb itself. Thus, particle verbs were included in the study.

2.2 Random forest model

To answer the question of which contextual cues are the most useful for deriving the meaning of the perception verb in the sentence, the full set of verbs was analysed using a random forest model (Breiman 2001; Tagliamonte & Baayen 2012). The random forest model is based on a set number of recursive partitioning trees. A single partitioning tree splits the given data into two so that observations with similar response variables are grouped together (Strobl et al. 2009). The splitting is continued until the algorithm reaches a preset condition. A random forest is a set of these trees. Using a random sampling method and comparing predicted values to observed values in the data, the random forest evaluates how important each predictor is (Tagliamonte & Baayen 2012). Thus, we end up with importance values assigned to all of the predictors – the higher the value, the more useful the predictor is in predicting the correct response variable.

The method has gained popularity in linguistics due to its relatively high success in dealing with disproportionate data. For example, when annotating a data sample with a large number of categorical variables, as is common in corpus-based language variation research, some of the predictor levels tend to have very few observations. This is not a problem with the tree-and-forest model, making it a very attractive method for linguists. However, it has been shown that the method gives preference to correlated predictor variables (Strobl et al. 2008).

In the present study, the random forest model was used to test which of the variables presented in Table 1 are the most significant predictors for the choice of perception verb in the corpus sample. In other words, it determines

the key contextual elements from which we could derive the meaning of the perception verb in the sentence.

3 Results

A random forest of 1,000 trees was used to test the contribution of the annotated variables. The forest was computed in R (R Core Team 2019) using the *party* package (Hothorn et al. 2006; Strobl et al. 2007; 2008). Table 2 shows the predicted vs. observed values in the model. The rows present the observed (actual) values and the columns present the predicted values. The table thus shows how many observations were predicted correctly for each verb and how many were mis-classified. For the mis-classifications, the table also shows which verb they were classified as. For example, *tundma* 'to feel' was classified as itself 394 cases out of 500, and one time as *maitset tundma* 'to taste' as well as *lõhna tundma* 'to smell'.

As can be seen from Table 2, the verb *maitset tundma* 'to taste' was the most mis-classified, with 200 classifications as *lõhna tundma* 'to smell', 14 classifications as *kuulma* 'to hear' and two classifications as *nägema* 'to see'. A rather clear division into two is observable from Table 2: the simple verbs vs. the compound verbs. Very rarely is one of the compound verbs *maitset tundma* 'to taste' or *lõhna tundma* 'to smell' classified as one of the simple verbs *nägema* 'to see', *kuulma* 'to hear' or *tundma* 'to feel', and vice versa. Using the *caret* package (Kuhn 2008) in R, the accuracy of the random forest model was calculated as 0.60.

The prediction accuracy of the model is rather low at 0.60 and this is mostly due to the fact that the verbs *maitset tundma* 'to taste' and *lõhna tundma* 'to smell' are not well predicted. Rather, they are almost equally predicted as each other. In addition, *nägema* 'to see' is also poorly predicted, with less than half of the observations predicted correctly.

Figure 1 shows the importance of all the variables included in the annotation of the corpus sample (see Table 1). Construction type is the most important factor for predicting the verb in the sentence (at 0.22). It is followed by semantic type of stimulus (at 0.12). Morphological person (at 0.05) and tense (at 0.03) are considerably less significant, followed by abstractness rating (at 0.03) and polarity (at 0.006), which has a minimal effect. Mode (at 0.001) and voice (at 0.0007) are not significant. The result mirrors results from Divjak (2015) in that for Russian perception verbs, the construction type

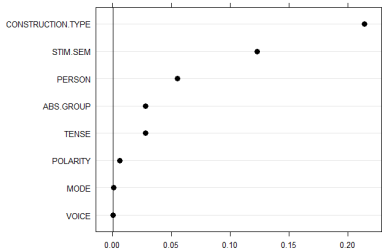

Figure 1. Variable importance plot (variables are shown on the y-axis, importance of the variables is shown on the x-axis)

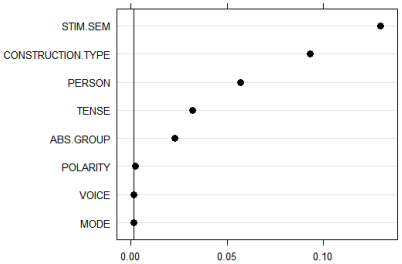

Figure 2. Variable importance plot for the model with *nägema*, *kuulma*, and *tundma* (variables are shown on the y-axis, importance of the variables is shown on the x-axis)

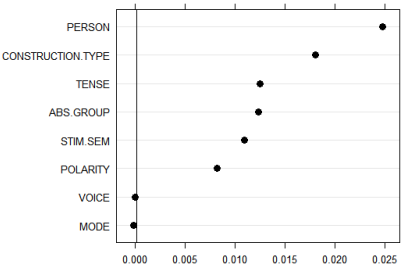

Figure 3. Variable importance plot for the model with *maitset tundma* and *lõhna tundma* (variables are shown on the y-axis, importance of the variables is shown on the x-axis)

Table 2. Predicted and observed values for the response variable in the random forest model

OBSERVED	PREDICTED				
	FEEL	HEAR	SEE	SMELL	TASTE
FEEL	**394**	47	57	1	1
HEAR	37	**345**	110	5	3
SEE	115	145	**235**	4	1
SMELL	2	30	0	**284**	184
TASTE	0	14	2	200	**201**

was the most important predictor as well. However, for Russian perception verbs, polarity was the second most significant contributor, and person was rather insignificant (Divjak 2015). No semantic predictors were included in Divjak (2015).

Because the verbs in the study separate into two distinct groups (simple verbs vs. compound verbs) according to their morphosyntax, it might be that the importance of the predictors is different for each group. To test this, the random forest model was also computed separately for each group, including only the verbs *nägema* 'to see', *kuulma* 'to hear', and *tundma* 'to feel' in one model, and the verbs *maitset tundma* 'to taste' and *lõhna tundma* 'to smell' in another model. The variable importance plots are presented in Figure 2 and Figure 3 respectively.

As is visible from Figures 2–3, separating the datasets changes the importance of the predictors considerably. When only the simple verbs are included in the model, the most significant variable is no longer construction type (0.09), but the semantic type of the stimulus (0.13). These are followed by person (0.05), tense (0.03), and abstractness rating (0.02). Polarity, voice and mode are not significant. For the compound verbs *maitset tundma* 'to taste' and *lõhna tundma* 'to smell', the most significant predictor is person (0.02), followed by construction type (0.01), tense (0.01), abstractness rating (0.01), semantic type of stimulus (0.01) and polarity (0.008). Voice and mode are insignificant. Thus, it seems that the contextual cues for deriving meaning are somewhat different for the two sets of verbs: for *nägema* 'to see', *kuulma* 'to hear' and *tundma* 'to feel' the key is the semantic type of stimulus, and for

maitset tundma 'to taste' and *lõhna tundma* 'to smell' the key is morphological person. Construction type is second most significant for both sets and takes clear precedence when the verbs are modelled together. Interestingly, the prediction accuracy of the models is not significantly higher than for the model that combines all the verbs. The accuracy is 0.61 for the compound verb set and 0.67 for the simple verb set.

In the next sections, the most important predictors are considered separately. As has recently been discussed by Gries (2019), conditional inference trees are not always easy to interpret, especially when dealing with a lot of variables and/or a large number of predictors. For the present study, a tree with the top three predictors was almost impossible to interpret because the visual representation (which is the basis for interpretation) could not sufficiently depict all the predictors. Thus, I present frequency tables for each of the predictors in order to give an overview of the variation for each variable in the corpus sample. The combination(s) of the predictors is discussed in the final section.

3.1 Construction type

Construction type is significantly more important than other predictors. It is also the predictor that exhibits the most variation in its levels, as seen from Table 1. However, it is important to note that some of the combinations of the verb and construction type are ungrammatical and it is to be expected that the frequency of these is zero.

As a language that has a rich morphosyntax, it is expected to see a broad variation in Estonian. For the simple verbs, the only mandatory element in the construction is the verb itself, since the verbs can also occur in intransitive phrases and Estonian is a pro-drop language where the verbal suffix alone can mark person. For the compound verbs, the nouns *maitse* 'taste' and *lõhn* 'smell' are also compulsory. Although the compound verbs cannot occur in syntactically intransitive phrases because of this, they can be used as semantically intransitive: in this case, the stimulus of the perception is not expressed, as exemplified in (4).

Table 3. Relative frequencies (%) of construction type per verb

	SEE	HEAR	FEEL	TASTE	SMELL
verb	4.4	14.4	1.8	0*	0*
verb as adjectival participle	0.6	0	11.6	0*	0*
verb + clause	21	20.4	9.8	0*	0*
verb + direct object	30.4	19.4	22	16.5	18.2
verb + direct object + adverb	15.2	13.6	6.4	1.9	3.6
verb + direct object + adjectival modifier	4	5.8	3	24.5	19.4
verb + direct object + nominal modifier	2.8	7.2	5.8	51.6	55.2
verb + direct object + nominal modifier + adjectival modifier	0	2	1.8	5.5	3.6
verb + direct object + adverbial	12	5.8	17	0*	0*
verb + adverbial	0	11	0	0*	0*
verb + reflexive particle + adjective phrase	0	0	12.2	0*	0*
verb + reflexive particle + noun phrase	0	0	2.6	0*	0*
phrasal verb + direct object	9.6	0.4	5.4	0*	0*
phrasal verb + adverbial	0	0	0.6	0*	0*
	100%	100%	100%	100%	100%

* grammatically impossible constructions

(4) *Sõrmeotsa-d on tundliku-d, ma tunnen maitseid ja lõhnu*
fingertip-PL be.3SG sensitive-PL I feel.1SG taste.PL.PART and smell.PL.PART
ja ega ma enamasti ei mõtle.
and PARTICLE I mostly NEG think.CONNEG

'(My) fingertips are sensitive, I can taste and smell and I'm mostly not thinking.'

Table 3 shows that the intransitive construction is mostly found for the verb *kuulma* 'to hear', where the verb is used intransitively in 14.4% of the sentences. This is because *kuulma* is also widely used as a discourse particle in the second singular and second plural person imperative. An example is provided in (5). The discourse particle has a function of attention grabbing

and/or addressing someone, and can also have a negative connotation. In the example, the verb is used for addressing someone, but with a judgemental undertone. From the present sample of verbs, *nägema* 'to see' can also occur as a discourse particle in the second person singular imperative (*näe*), and also functions as an attention-grabbing device. Perception verbs being used as conversational particles or grammaticalizing into conversational particles has been observed in a number of languages by San Roque et al. (2018: 388–389).

(5) *kuule* *mees,* *mis* *sa* *kogu* *aeg* *selle* *Põhja-Korea*
hear.2SG.IMPR man what you whole time this.GEN North-Korea.GEN
kallal *nori-d?*
at pick-2SG

'Hey dude, why are you always picking on North-Korea?'

The corpus sample shows that all of the perception verbs can form constructions with the stimulus as one of the construction participants. The morphological form of the stimulus varies. For *nägema* 'to see', *kuulma* 'to hear' and *tundma* 'to feel', the stimulus mostly occurs in cases that are considered syntactic object cases in Estonian: genitive and partitive. A very clear pattern emerges when the stimulus is realised as an adverbial as opposed to an object. Estonian grammar distinguishes between compulsory and not compulsory adverbials (Erelt & Metslang 2017: 300), and in the present study the construction "verb + adverbial" represents only the cases of a compulsory adverbial. Only *kuulma* 'to hear' occurs in this construction, and it makes up 11% of all the construction types for *kuulma* 'to hear'. The stimulus in this case is realised in the elative case as in example (6). When the stimulus is in the genitive or partitive case in *kuulma*-sentences, the construction retains its perception experience meaning. However, the elative-stimulus construction's meaning shifts to "reported knowledge" as exemplified in example (6).

(6) *Näiteks* *Saksamaa-l,* *mille* *võõrtööliste* *probleemi-de-st*
for.example Germany-ADE which.GEN foreign.worker.PL.GEN problem-PL-ELA
me *alatasa* *kuuleme,* *on* *vaid* *kaheksa* *protsenti* *immigrante.*
we always hear.1PL be.3SG only eight percent.PART immigrant.PL.PART

'For example, we are constantly hearing about problems with the foreign workforce in Germany, but the percentage of immigrants there is only eight.'

For other verbs, construction participants that occur in locative or other semantic cases lexicalise some other event participants like the theme or location as exemplified in (7).

(7) *Egon Tintse foto-l näeme hetke mulluste-st*
Egon Tintse photo-ADE see.1PL moment.PART yesteryear.PL-ELA
omavaheliste-st heitluste-st.
private.PL-ELA fight.PL-ELA

'In the photo by Egon Tintse, we can see a moment from the fight from last year.'

Another possibility of coding the perception stimulus in the sentence is showcased by *maitset tundma* 'to taste' and *lõhna tundma* 'to smell'. A very clear pattern emerges when the stimulus is coded as a noun phrase, adjective phrase, or both. The modifiers lexicalise the source or the quality of the gustatory or olfactory experience. The object can be modified by an NP or an AP or both for both verbs. Some examples are provided in (8–10). Example (8) is also an example of a passive construction that is common to perception verbs (and some mental verbs). This construction is formed by combining the perception verb in the *da*-infinitive with the verb *olema* 'to be' (Erelt & Metslang 2017: 223–224). In this construction, the nouns *maitse* 'taste' and *lõhn* 'smell' can be in the nominative case in addition to the prototypical object case partitive.

(8) *Uuesti ei proovi-ks vist, riivi-si-n küll jogurti sisse aga*
again NEG try-COND maybe grate-PST-1SG enough yoghurt.GEN into but
ikka oli seda jubeda-t maitse-t tunda.
still be.3SG.PST this.PART awful-PART taste-PART feel.INF1

'I do not think I would try it again, I grated it into yoghurt but it still tasted awful.'

(9) *Tekib illusioon, nagu või-ks ninasõõrme-te-s tunda balsameeritu*
arise.3SG illusion like can-COND nostril-PL-INE feel.INF1 embalmed.GEN
lõhna.
scent.PART

'An illusion is created as if one can smell the embalmed in your nose.'

(10) *Arvuti asu-b mul just mesilaste pesitsuspaiga all*
computer be.located-3SG I.ADE PARTICLE bee.PL.GEN nest.GEN below
ning magusa-t mee lõhna ikka tundsin.
and sweet-PART honey.GEN smell.PART PARTICLE feel.1SG.PST

'My computer is located exactly under the bees' nest so I could really smell the sweet honey.'

Some constructions are comprised only of the verb and its morphological form. For example, when the verb occurs as an adjectival participle, it is

almost certainly *tundma* ‘to feel’ as in example (11). The morphological passive past participle in Estonian, when occurring before a noun phrase, acts more like an adjective, as it describes some characteristic of the noun it modifies. *Tundma* ‘to feel’ in this form is near-synonymous with ‘famous’. This construction makes up 11.6% of all the constructions for *tundma* ‘to feel’. *Nägema* ‘to see’ also occurs in this construction in the sample, but for *nägema* it only makes up 0.6% of all the cases. *Tundma* ‘to feel’ is also predicted via a reflexive construction (*tunnen ennast* feel.1SG oneself.PART) – it is the only verb that occurs in this construction in the sample (14.8% of all the constructions of *tundma*). An example of the reflexive construction is provided in (12).

(11) *Seda on öelnud tuntud teoloog ja usundiloolane Karl Barth.*
this.PART be.3SG say.APP feel.PPP theologian and religion.specialist Karl Barth

‘This has been said by the famous theologian and religion specialist Karl Barth.’

(12) *küsi ta-lt midagi, tunne ennast mugavalt, sest see on eesmärk*
ask.2SG.IMP (s)he-ABL something feel.2SG.IMP oneself.PART comfortably because this be.3SG goal

‘ask him/her something, be relaxed, because that is the goal’

Only the simple verbs *nägema* ‘to see’, *kuulma* ‘to hear’ and *tundma* ‘to feel’ can also have the stimulus of the perception coded in the form of a clause, making up 21%, 20% and 9.8% of all the constructions respectively. This is inevitable due to the fact that the “object slot” is already filled for *maitset tundma* ‘to taste’ and *lõhna tundma* ‘to smell’. An example of the clausal stimulus is provided in (13).

(13) *Olen ise olnud saksa keele õpetaja ja näinud, kuidas lastel on raske enda motiveerimise-ga hakkama saada.*
be.1SG self be.APP German language teacher and see.APP how child.PL.ADE be.3PL difficult self.GEN motivating-COM start.INF2 become.INF1

‘I have been a German teacher myself and I have seen how difficult it is for the children to motivate themselves.’

The simple verbs can combine with a number or particles, making the meaning of the verb not (always) transparent. In the present study, particle verbs were not excluded from the data in the initial manual-checking phase, since the goal was to explore all kinds of different structures available from the corpus sample. *Tundma* 'to feel' occurs as a particle verb in 6% of all the construction types, *nägema* 'to see' in 9.6% of the cases and *kuulma* 'to hear' in 0.4% of the cases. In example (14), *nägema* combines with the particle *ette* 'to the front'. The particle verb *ette nägema* is polysemous on its own; it can mean 'prescribe' or 'foresee', thereby also echoing the locative meaning of the particle.

(14) *õpetaja-le* *nähakse* *ette* *üldtööaja-le*
teacher-ALL see.IMPRS.PRS front.PARTICLE general.work.time-ALL
vastav *ametipalk*
corresponding occupation.wage
'a wage that corresponds to the general working time is assigned to the teacher'

Other particle verbs in the sample were *pealt nägema* 'witness' (lit. 'from the top see'), *läbi nägema* 'see through (someone or something)', and *ära nägema* 'see' (*ära* is an aspectual marker in Estonian which denotes completeness of a process).[8] *Tundma* 'to feel' creates a particle verb with the particles *ära* and *kaasa* 'with'. *Ära tundma* means 'recognise' and *kaasa tundma* means 'sympathise with someone'; thus, these are polysemous meanings of *tundma* 'to feel' that are realised only as a specific morphosyntactic construction. There are only a few examples of particle verbs with *kuulma* and they are always with the particle *pealt* 'from the top' – the meaning shifts to 'overhear'.

3.2 Semantic type of stimulus

Variation of the semantics of the stimuli of the perception verbs speak to the polysemy of the verbs. As shown in Table 4, there are three types of stimuli that occur with all of the five perception verbs: perceivable stimuli, concrete stimuli and abstract stimuli. However, within these groups there is considerable variation in which verbs occur with which semantic types of stimuli.

Most notable in this respect is *tundma* 'to feel', which has a very minimal amount of "sense" stimuli (2.2%), but a very high proportion of "abstract"

[8] See Metslang (2001) for further discussion about the particle *ära* in Estonian.

Table 4. Relative frequencies (%) of stimulus types per verb

	SEE	HEAR	FEEL	TASTE	SMELL
abstract	11.9	3.6	41.8	10.3	4.4
ambivalent	7.4	5	6.1	0	0
concrete	21.5	26.1	22.1	1.7	5.8
person	6.1	2.3	16.6	0	0
sense	26	40.3	2.2	88	89.8
situation/event	27.2	22.8	11.2	0	0
	100%	100%	100%	100%	100%

stimuli (41.8%), examples are provided in (15) ("sense" stimulus) and (16) ("abstract" stimulus). Thus, we can say that when a perception verb occurs in a sentence with an abstract stimulus, it is most likely *tundma* 'to feel', and when the stimulus is a perceivable entity, it is the least likely that the verb in the sentence is *tundma*.

(15) *Andres mäleta-b, et vahetult pärast õnnetust ta valu*
Andres remember-3SG that immediately after accident.PART he pain.GEN
ei tundnud-ki.
NEG feel.APP-CLITIC

'Andres remembers that immediately after the accident he did not even feel any pain.'

(16) *Miks me tunneme hirmu, kui me välju-me turvatsooni-st?*
why we feel.1PL fear.PART when we exit-1PL safe.zone-ELA

'Why do we feel fear when we step out of the safety zone?'

The stimulus type "person" mostly occurs with *tundma* 'to feel' and makes up 16.6% of all the stimuli for the verb. For *nägema* 'to see' the percentage of "person" stimuli is 6.13%, and for *kuulma* 'to hear' the proportion is only 2.3%. This reflects the fact that both *nägema* 'to see' and *tundma* 'to feel' have polysemous meanings when occurring with a "person" type stimulus. 'Seeing a person' in Estonian implies social contact with said person, not only visual perception (Proos 2019) (e.g. *Pole sind ammu näinud!* 'I haven't seen you in a long time!'). 'Feeling a person' is also a type of personal contact and the meaning in Estonian is 'knowing someone thoroughly' (Proos 2020b)

(e.g. *Tunnen Jaani lapsest saati* 'I have known Jaan since we were children'). *Maitset tundma* 'to taste' and *lõhna tundma* 'to smell' do not occur with this type of stimulus.

The latter also do not occur with the type "situation/event". The "situation/event" type is mostly lexicalised by a clausal object, so this correlates with the construction type pattern, which was also not observed for *maitset tundma* and *lõhna tundma*. However, *maitset tundma* 'to taste' and *lõhna tundma* 'to smell' most frequently take a "sense" stimulus out of all the verbs at 88% and 89.8% respectively. These differences in co-occurrences are reflected in Figure 2, where the semantic type of stimulus was shown as the most significant contributor. The stimulus types "sense" and "abstract" are especially informative as to which verb occurs in the sentence.

3.3 Person

As is visible from Figure 1, (morphological) person is a much less significant predictor than construction type and semantic type of stimulus. From Table 5 we can also see a rather balanced distribution, i.e. for all the five verbs, the different morphological forms occur in similar proportions.

The significance of this predictor comes from a few proportions that stand out. For example, *tundma* 'to feel' occurs as the passive past participle in 16.5% of the cases: the proportion is significantly lower for other verbs. Note that this result is also reflected in the construction type variable – the construction "verb used as adjectival participle" corresponds to "passive past participle" on the morphological level. This is due to the fact that the active past participle form of *tundma* 'to feel' (*tuntud*) is a polysemous meaning of the verb. *Tuntud* means 'famous' in Estonian, e.g. *tuntud näitleja* 'a famous actor' (lit. 'felt actor'). The proportion of infinitives is rather large for all of the verbs, even reaching 59.1% for *maitset tundma* 'to taste'. The exception here is *tundma* 'to feel' with only 10.8% of the sentences having the verb in the infinitive form.

Figure 3 showed that when only *maitset tundma* 'to taste' and *lõhna tundma* 'to smell' are entered into the model, morphological person is the most informative predictor as to which verb is present in the sentence. From Table 5 we can see that three variable levels have a noticeable difference between 'smell' and 'taste' co-occurrences. *Lõhna tundma* 'to smell' occurs in the first person singular 7.2% more frequently and in the third person singular 8.6% more frequently. *Maitset tundma* 'to taste' occurs in the infinitive 14.7% more frequently.

Table 5. Relative frequencies (%) of morphological person per verb

	SEE	HEAR	FEEL	TASTE	SMELL
1SG	16.73	25.10	16.53	9.11	16.29
1PL	8.67	7.11	3.27	3.45	3.26
2SG	2.42	6.69	6.33	4.93	4.07
2PL	2.22	6.69	3.27	2.46	2.44
3SG	15.52	15.90	26.94	11.82	20.37
3PL	6.05	5.44	11.22	8.87	7.94
infinitive	38.51	29.71	10.82	59.11	44.40
active past participle	0.40	0.21	0.20	0.00	0.00
passive past participle	6.05	2.72	16.53	0.00	0.41
impersonal present	2.22	0.21	3.67	0.00	0.61
impersonal past	1.21	0.21	1.22	0.25	0.20
	100%	100%	100%	100%	100%

4 Discussion

The present study looked at Estonian experience perception verbs and their usage patterns in a corpus sample. The aims of the study were to:

1. explore how verbs of the same class (experience perception verbs) but of different morphosyntax (simple vs compound verbs) behave in a corpus sample;

2. determine which contextual elements would best allow the speaker to derive the meaning of the perception verb in the sentence.

All of the verbs can occur with a stimulus that encodes something that can be perceived via the respective sense. This is the physical or perceptual meaning that all perception verbs have. The realisation pattern of the stimulus, however, can be different, as already illustrated in examples (1–3). As described by Croft (2012: 233), all of the perception verbs were shown to include a stimulus as one of the possible participants in their event construal, but all of the verbs also exhibited the volitionality of including the stimulus, i.e. they could occur in intransitive phrases. Thus, we do see some patterns

that unify the class of verbs as one. Considerable differences between the individual verbs were, however, also found.

As illustrated by Tables 3–5 as well as by Figures 2–3, the experience perception verbs in Estonian split into two in regard to their corpus behaviour. On the one hand, the simple verbs *nägema* 'to see', *kuulma* 'to hear', and *tundma* 'to feel' exhibit behaviour that is similar in many ways, while the compound verbs *maitset tundma* 'to taste' and *lõhna tundma* 'to smell' form another group that exhibits considerable similarities. This split has two reasons: the morphosyntactic composition of the verbs itself sets limits to the possible variation, and in addition, the different degree of polysemy amongst the two types of verbs also plays a role.

Although this study did not explicitly consider polysemy as one of the factors for reasons outlined in § 1, polysemy is very tightly connected to the variables that were included in the study. For example, polysemy is the reason why, for some verbs, the semantic type of stimulus is often not a perceivable, physical one, but rather falls into one of the other categories. For example, only 2.23% of stimuli are perceivable for the verb *tundma* 'to feel'. At the same time, *tundma* 'to feel' occurred with a lot of stimuli from the types "abstract" and "person". As is pointed out in § 3, "abstract" and "person" types of stimuli are representative of different polysemous meanings of *tundma* 'to feel'. *Tundma* 'to feel' is a highly polysemous verb; in fact, Proos (2020b) has analysed it as more like a general proximal perception verb as opposed to a prototypical tactile experience perception verb. This is reflected in both the construction types that *tundma* occurs in as well as in the variety of stimuli it can occur with. *Tundma* 'to feel' also exhibits a lot of morphological variation as showcased by Table 5, allowing it to form even more constructions.

Similarly to *tundma* 'to feel', *nägema* 'to see' is also very polysemous (Proos 2019) and thus exhibits a large degree of variation in both the semantic types of stimuli it occurs with and morphosyntactic patterns. Similarly to *tundma* 'to feel', when *nägema* 'to see' occurs with the stimulus type "person", it is representative of a specific polysmous meaning, namely 'to meet a person'. However, contrary to *tundma* 'to feel', *nägema* 'to see' occurs a lot with perceivable stimuli (26% of the stimuli). This result echoes the cross-linguistically attested prominence of vision as the sense that is talked about the most (San Roque et al. 2014).

On the other side of the polysemy spectrum are the two compound verbs *maitset tundma* 'to taste' and *lõhna tundma* 'to smell'. Although the limited

variation due to their morphosyntactic composition is evident, this need not necessarily limit the semantic variation a compound verb exhibits. Yet, the compound verbs *maitset tundma* 'to taste' and *lõhna tundma* 'to smell' are the least semantically varied as well, as is illustrated by the low level of variation in the semantic types of stimuli. The verbs are not very polysemous; the only extended meanings they exhibit are fixed expressions like *tunnen võidu maitset* 'I (can) taste the victory' or *tundsin tehingu juures raha lõhna* 'I could smell money on the deal'. Since their inherent semantic potential for extension into fields other than perception is limited, these verbs cannot occur in all the constructions that *nägema* 'to see', *kuulma* 'to hear' and *tundma* 'to feel' can occur in. Thus, the richness of a verb's variation is influenced both by the constructions it occurs in as well as the semantic content of that verb – these two sides are intertwined without necessarily establishing a causal link between them.

Polysemy also plays a role in determining which contextual elements would best help the speaker in deriving the meaning of the perception verb in the sentence. The results show that the most significant predictor amongst the variables is construction type. In some cases, a construction type realises a certain polysemous meaning of the verb. For example, (17) is an example of a polysemous meaning of *tundma* 'to feel'. The reflexive construction is indicative of a specific meaning of *tundma* 'to feel' – 'feeling like someone/something; being in a position'. Similarly, when *kuulma* 'to hear' occurs with a stimulus in the elative case, it represents the polysemous meaning of *kuulma* – hearsay or reported knowledge, as exemplified by (18)[9] (this meaning is also reported by Vanhove (2008: 348–349) as 'learn' and 'know the story of' in English).

(17) *Kas tunned ennast juhi-na, kes on sattu-nud aja*
QPART feel.2SG oneself.PART leader-ESS who be.3SG end.up-PPP time.GEN
hammasrataste vahele?
cogwheel.PL.GEN between.ALL

'Do you feel like a leader who got in the way of time?'

[9] Example (6) is repeated here as (18).

(18) *Näiteks* *Saksamaa-l,* *mille* *võõrtööliste*
for.example Germany-ADE which.GEN foreign.worker.PL.GEN
probleemi-de-st *me* *alatasa* *kuuleme,* *on* *vaid* *kaheksa* *protsenti*
problem-PL-ELA we always hear.1PL be.3SG only eight percent.PART
immigrante.
immigrant.PL.PART

'For example, we are constantly hearing about problems with the foreign workforce in Germany, but the percentage of immigrants there is only eight.'

The results also indicate that the semantic type of stimulus could be an important predictor of the meaning of the perception verb in the sentence. Interestingly, here we can observe a pattern that unifies all of the verbs except *kuulma* 'to hear'. When the stimulus is something abstract, it almost always hints to some sort of emotional meaning of the perception verbs. The pattern is, however, limited in the cases of *nägema* 'to see', *maitset tundma* 'to taste' and *lõhna tundma* 'to smell'. These verbs only occur with a limited number of different emotion-related stimuli and rather infrequently. Examples (19–21) illustrate this pattern.

(19) *Nägime* *kurja* *vaeva,* *aga* *viimase-l* *minuti-l*
see.1PL.PST evil.PART pain.PART but last-ADE minute-ADE
jaota-ti *kursus* *teatri-te* *vahel* *laiali.*
distribute-IMPRS.PST course theater-PL.GEN among spread.out

'We really tried, but at the last minute, the course was dispersed between different theatres.'

(20) *Ma* *mõistsin,* *et* *kui* *oskan* *kaotada,* *siis*
I understand.1SG.PST that if know.how.1SG lose.INF1 then
oskan *tunda* *ja* *nautida* *võidu* *magusa-t* *maitse-t.*
know.how.1SG feel.INF1 and enjoy.INF1 victory.GEN sweet-PART taste-PART

'I realised that if I know how to lose, I can taste and enjoy the sweetness of victory.'

(21) *Ärevuse* *lõhna* *on* *lausa* *tunda,* *sest* *viimase-d*
anxiety.GEN smell.PART be.3SG even feel.INF1 because last-PL
tulemuse-d *pole* *kõigi-le* *veel* *teada.*
result-PL be.3SG.NEG everyone-ALL yet know.INF1

'You can even smell the anxiety, because the final results are not known yet.'

Tundma ‘to feel’ however is rather productive in this construction. This is also tied to the polysemy of the verbs. *Tundma* ‘to feel’ as a proximal perception verb also covers feeling emotions and thus is very productive with various sorts of emotion-related stimuli. *Nägema*, *maitset tundma* and *lõhna tundma* in this construction are idiomatic, which is why the number of different stimuli that they can occur with is so limited. However, it is quite striking that four out of five experience perception verbs in Estonian have the potential to express emotional experiences. Perhaps this hints to the pervasive ties that exist between perception and cognition.

It also seems that the semantic type of stimulus “concrete” is predictive of a polysemous meaning that expresses some type of comprehension or analysis of things as exemplified by examples (22–24). In example (22), *nägema* ‘to see’ has the polysemous meaning of ‘understand’ and expresses comprehending and analysing the results of some kind of sales strategy. In example (23), however, *tundma* has the polysemous meaning of ‘having extensive knowledge of or skill in a field’. In (24), the person does hear the diagnosis, but it is more than an auditory experience: a level on comprehension and understanding is also inherently encoded in the construction. Thus, (24) can also be considered as a case of a metonymy relationship. However, the “concrete” type of stimulus is very infrequent with the compound verbs; thus, we see again a type of split between the simple verbs and the compound verbs.

(22) *siis on seda ka keeruline müüa ja näiteks praegu*
then be.3SG this.PART also difficult sell.INF1 and for.example now
meie näeme tulemusi selle-st alles septembri-s, sest
we see.1PL result.PL.PART this-ELA not.before september-INE because
kaks kuud kõik puhkavad siin
two month.PART everyone rest.3PL here

‘so it is difficult to sell and at the moment, for example, we will see the results of this only in September, since everyone is on vacation here for two months’

(23) *Marsi sisemust tuntakse ainult pinna-lt saa-dud*
Mars.GEN inside.PART feel.IMPRS.PRS only surface-ABL get-PPP
andme-te ja planeedi üldstatistika kaudu.
data-PL.PART and planet.GEN general.statistics.GEN through

‘All the information about the insides of Mars is based off of data from the surface and general statistics about the planet.’

(24) *kuid tema ütles diagnoosi kuuldes hoopis ”vaat kui*
but (s)he say.3SG.PST diagnosis.PART hear.GER quite look.PARTICLE how
huvitav!”
interesting

‘but instead, when (s)he heard the diagnosis, (s)he said “well that’s interesting!” ’

The results show that there are some usage patterns that unify the experience perception verb class as a whole, but there are more that are predictive of only one perception verb or a set of similar verbs. The notion of usage patterns combines function both on the level of form and meaning. From the results, we see that on the one hand, the specific verb’s own characteristics, e.g. its semantic potential, are important for modelling the usage pattern variation, and on the other hand, the constructions we see emerge on the syntax level also bring variation into the usage patterns. There is an innate link between meaning and form, and variation on the morphosyntactic level is also tied to the variation on the semantic level.

The results also show that it is possible to see the variation of semantic function within one verb, i.e. polysemy, by looking at the semantic and morphosyntactic characteristics occurring with the verb in the sentence. Some polysemous meanings of the verbs are construction-specific in that the form and meaning relationship has conventionalised to the extent that the particular pairing is understood as a separate meaning. Although this approach does not allow the composing of a conclusive list of polysemous meanings, it reflects how meaning variation can be an inherent part of general usage pattern variation.

In regards to the importance of the construction type, the results of this study reflect results from Divjak (2015) as well as Janda & Solovyev (2009). Thus, this study confirms yet again how important the role of form is to constructing meaning. Especially in languages with rich morphology, the information coded into the form of the verb or construction as a whole is already considerably informative about the meaning of a verb in the sentence. This is not to say that the language user makes calculations on the basis of co-occurrence information; rather, this hints towards there being a certain amount of co-occurrence and context information that the language user applies in their everyday language comprehension and production.

5 Conclusion

This study looked at five Estonian experience perception verbs: *nägema* 'to see', *kuulma* 'to hear', *tundma* 'to feel', *maitset tundma* 'to taste' and *lõhna tundma* 'to smell'. The aim of the study was to analyse the usage patterns of these verbs in a corpus sample, and to determine which contextual elements could be the most informative about which perception verb is present in the sentence. It was expected that the verbs share some usage patterns because they belong to the same verb class of perception verbs; however, they should also exhibit considerable dissimilarities in their usage patterns because of their differing morposyntax and degree of polysemy.

A random forest model was constructed, which showed the high importance of construction type in predicting the verb in the sentence, and, to a lesser extent, the importance of the semantic type of the stimulus. It was shown how morphosyntactic variation goes hand in hand with semantic variation, and how the usage patterns of the perception verbs combine these two sides.

Acknowledgements

This study has been supported by a research grant from the Estonian Research Council (PUT1358 "The Making and Breaking of Models: Experimentally Validating Classification Models in Linguistics").

Abbreviations

1, 2, 3	person
ABL	ablative
ADE	adessive
ALL	allative
APP	active past participle
CLITIC	clitic
COM	comitative
COMP	comparative
COND	conditional
CONNEG	connegative

ELA	elative
ESS	essive
GEN	genitive
GER	gerund
ILL	illative
IMPR	imperative
IMPRS	impersonal
INE	inessive
INF1	1st infinitive
INF2	2nd infinitive
NEG	negation
NOM	nominative
PART	partitive
PARTICLE	particle
PL	plural
PPP	passive past participle
PRS	present
PST	simple past
QPART	question particle
QUOT	quotative
SG	singular

References

Aedmaa, Eleri. 2019. *Detecting compositionality of Estonian particle verbs with statistical and linguistic methods*. Tartu: University of Tartu. (Doctoral dissertation).

Berez, Andrea L. & Gries, Stefan T. 2008. In defense of corpus-based methods: A behavioral profile analysis of polysemous get in English. In Moran, Steven & Tanner, Darren S. & Scanlon, Michael (eds.), *Proceedings of the 24th Northwest Linguistics Conference*, 157–166. Seattle, WA: University of Washington Working Papers in Linguistics.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1). 5–32.

Bresnan, Joan & Cueni, Anna & Nikitina, Tatiana & Baayen, R. Harald. 2007. Predicting the dative alternation. In Bouma, Gerlouf & Krämer, Irene & Zwarts, Joost (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan & Ford, Marilyn. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213.

Croft, William. 2012. *Verbs: Aspect and causal structure*. New York: Oxford University Press.

Divjak, Dagmar. 2010. *Structuring the lexicon: A clustered model for near-synonymy*. Berlin: Walter de Gruyter.

— 2015. Exploring the grammar of perception: A case study using data from Russian. *Functions of Language* 22(1). 44–68.

Divjak, Dagmar & Fieller, Nick. 2014. Cluster analysis: Finding structure in linguistic data. In Glynn, Dylan & Robinson, Justyna A. (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*, 405–441. (Human Cognitive Processing 43). Amsterdam: John Benjamins.

Divjak, Dagmar & Gries, Stefan T. 2006. Ways of trying in Russian: Clustering behavioural profiles. *Corpus Linguistics and Linguistic Theory* 2(1). 23–60.

Erelt, Mati & Metslang, Helle (eds.). 2017. *Eesti keele süntaks* [Estonian syntax]. (Eesti Keele Varamu III). Tartu: Tartu Ülikool kirjastus.

etTenTen. 2013. *Estonian web corpus*. Center of Estonian Language Resources. (https://www.keeleveeb.ee/dict/corpus/ettenten/). (Accessed 2019-04-01).

Evans, Nicholas & Wilkins, David. 2000. In the mind's ear: The semantic extensions of perception verbs in Australian languages. *Language* 76(3). 546–592.

Franco, Karlien & Geeraerts, Dirk & Speelman, Dirk & Van Hout, Roeland. 2019. Concept characteristics and variation in lexical diversity in two Dutch dialect areas. *Cognitive Linguistics* 30(1). 205–242.

Glynn, Dylan. 2014. The many uses of *run*: Corpus methods and socio-cognitive semantics. In Glynn, Dylan & Robinson, Justyna A. (eds.), *Corpus methods for semantics: Quantitive studies in polysemy and synonymy*, 117–144. (Human Cognitive Processing 43). Amsterdam: John Benjamins.

— 2016. Quantifying polysemy: Corpus methodology for prototype theory. *Folia Linguistica* 50(2). 413–447.

Grafmiller, Jason & Szmrecsanyi, Benedikt. 2018. Mapping out particle placement in Englishes around the world: A study in comparative sociolinguistic analysis. *Language Variation and Change* 30(3). 385–412.

Gries, Stefan T. 2006. Corpus-based methods and cognitive semantics: The many senses of *to run*. In Gries, Stefan T. & Stefanowitsch, Anatol (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 57–99. Berlin: Mouton de Gruyter.

— 2019. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*. (Ahead of print publication). DOI: 10.1515/cllt-2018-0078.

Harris, Zellig S. 1954. Distributional structure. *WORD* 10(2–3). 146–162.

Hothorn, Torsten & Buehlmann, Peter & Dudoit, Sandrine & Molinaro, Annette & Van Der Laan, Mark J. 2006. Survival ensembles. *Biostatistics* 7(3). 355–373.

Ibarretxe-Antuñano, Iraide. 2008. Vision metaphors for the intellect: Are they really cross-linguistic? *Atlantis: Journal of the Association of Anglo-American Studies* 30(1). 15–33.

Janda, Laura A. & Solovyev, Valery D. 2009. What constructional profiles reveal about synonymy: A case study of Russian words for sadness and happiness. *Cognitive Linguistics* 20(2). 295–324.

Klavan, Jane. 2020. Pitting corpus-based classification models against each other: A case study for predicting constructional choice in written Estonian. *Corpus Linguistics and Linguistic Theory* 16(2). 363–391.

Köper, Maximilian & Schulte im Walde, Sabine. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350,000 German lemmas. In Goggi, Sara & Mazo, Hélène & Kosem, Iztok & Vintar, Špela & Krek, Simon (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2595–2598. Portorož: European Language Resources Association (ELRA).

Kuhn, Max. 2008. Building predictive models in R using the **caret** package. *Journal of Statistical Software* 28(5).

Langemets, Margit & Tiits, Mai & Valdre, Tiia & Veskis, Leide & Viks, Ülle & Voll, Piret (eds.). 2009. *Eesti keele seletav sõnaraamat* [Explanatory dictionary of Estonian]. Tallinn: Eesti Keele Sihtasutus.

Metslang, Helle. 2001. On the developments of the Estonian aspect: The verbal particle *ära*. In Dahl, Östen & Koptjevskaja-Tamm, Maria (eds.), *Circum-Baltic languages, vol. II: Grammar and typology*, 443–479. (Studies in Language Companion Series 55). Amsterdam: John Benjamins.

Peirsman, Yves & Geeraerts, Dirk & Speelman, Dirk. 2010. The automatic identification of lexical variation between language varieties. *Natural Language Engineering* 16(4). 469–491.

Proos, Mariann. 2019. Polysemy of the Estonian perception verb "nägema" 'to see.' In Speed, Laura J. & O'Meara, Carolyn & San Roque, Lila & Majid, Asifa (eds.), *Perception metaphors*, 231–252. (Converging Evidence in Language and Communication Research 19). Amsterdam: John Benjamins.

— 2020a. *Data & script for the article "Usage patterns of Estonian experience perception verbs: A random forest approach"*. Open Science Framework. (https://doi.org/10.17605/OSF.IO/JSHVF). (Accessed 2020-04-14).

— 2020b. Feeling your neighbour: An experimental approach to the polysemy of *tundma* 'to feel' in Estonian. *Language and Cognition* 12(2). 282–309.

R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. (https://www.R-project.org/). (Accessed 2019-04-01).

Rätsep, Huno. 1978. *Eesti keele lihtlausete tüübid* [Estonian simple clause types]. Tallinn: Valgus.

San Roque, Lila & Kendrick, Kobin H. & Norcliffe, Elisabeth & Brown, Penelope & Defina, Rebecca & Dingemanse, Mark & Dirksmeyer, Tyko & Enfield, N. J. & Floyd, Simeon & Hammond, Jeremy & Rossi, Giovanni & Tufvesson, Sylvia & van Putten, Saskia & Majid, Asifa. 2014. Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies. *Cognitive Linguistics* 26(1). 31–60.

San Roque, Lila & Kendrick, Kobin H. & Norcliffe, Elisabeth & Majid, Asifa. 2018. Universal meaning extensions of perception verbs are grounded in interaction. *Cognitive Linguistics* 29(3). 371–406.

Strobl, Carolin & Boulesteix, Anne-Laure & Kneib, Thomas & Augustin, Thomas & Zeileis, Achim. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9(307).

Strobl, Carolin & Boulesteix, Anne-Laure & Zeileis, Achim & Hothorn, Torsten. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8(25).

Strobl, Carolin & Malley, James & Tutz, Gerhard. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4). 323–348.

Sweetser, Eve E. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. (Cambridge Studies in Linguistics 54). Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt & Grafmiller, Jason & Rosseel, Laura. 2019. Variation-based distance and similarity modeling: A case study in world Englishes. *Frontiers in Artificial Intelligence* 2(23).

Tagliamonte, Sali A. & Baayen, R. Harald. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.

Vanhove, Martine. 2008. Semantic associations between sensory modalities, prehension and mental perceptions: A crosslinguistic perspective. In Vanhove, Martine (ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 341–370. (Studies in Language Companion Series 106). Amsterdam: John Benjamins.

Viberg, Åke. 1984. The verbs of perception: A typological study. *Linguistics* 21. 123–162.

— 2015. Sensation, perception and cognition: Swedish in a typological-contrastive perspective. *Functions of Language* 22(1). 96–131.

Contact information:

Mariann Proos
Institute of Estonian and General Linguistics
University of Tartu
Jakobi 2-446
51007 Tartu
Estonia
e-mail: mariann(dot)proos(at)ut(dot)ee