

Volume 32:2019

sky

SKY Journal of Linguistics

Editors:

Johanna Isosävi, Antti Kamunen, Aleksi Mäkilähde,
Santeri Palviainen, Tomi Visakko

SKY Journal of Linguistics is published by the Linguistic Association of Finland (one issue per year).

Notes for Contributors

Policy: *SKY Journal of Linguistics* welcomes unpublished original works from authors of all nationalities and theoretical persuasions. Every manuscript is reviewed by at least two anonymous referees. In addition to full-length articles, the journal also accepts short (3–9 pages) ‘squibs’ as well as book reviews.

Language of Publication: Contributions should be written in English, French, German, Finnish, or Swedish. If the article is not written in the native language of the author, the language should be checked by a professional in that language.

Style Sheet: *SKY Journal of Linguistics* follows the Generic Style Rules for Linguistics, with complementary house rules, which can be found at <http://www.linguistics.fi/skystyle.shtml>.

Abstracts: Abstracts of the published papers are included in *Linguistics Abstracts* and *Cambridge Scientific Abstracts*. The published papers are included in *EBSCO Communication & Mass Media Complete*. *SKY JoL* is also indexed in the *MLA Bibliography*.

Editors’ Addresses (2019):

Johanna Isosävi, Department of Modern Languages, P. O. Box 24 (Unioninkatu 40), FI-00014 University of Helsinki, Finland

Antti Kamunen, English Philology, Faculty of Humanities, P. O. Box 1000, FI-90014 University of Oulu, Finland

Aleksi Mäkilähde, Department of English, FI-20014 University of Turku, Finland

Santeri Palviainen, Finnish Language, Faculty of Humanities, P. O. Box 1000 FI-90014 University of Oulu, Finland

Tomi Visakko, Department of Finnish, Finno-Ugrian and Scandinavian Studies, P. O. Box 3, FI-00014 University of Helsinki, Finland

Editors’ E-mail: [sky-journal\(at\)helsinki\(dot\)fi](mailto:sky-journal(at)helsinki(dot)fi)

Publisher:

The Linguistic Association of Finland
c/o Tommi Alho
Åbo Akademi University
Department of English Language and Literature
Tehtaankatu 2
FI-20500 Turku
Finland

<http://www.linguistics.fi>, <http://www.linguistics.fi/skyjol.shtml>

The Linguistic Association of Finland was founded in 1977 to promote linguistic research in Finland by offering a forum for the discussion and dissemination of research in linguistics, both in Finland and abroad. Membership is open to anyone interested in linguistics. The membership fee in 2019 was EUR 25 (EUR 15 for students and unemployed members).

SKY Journal of Linguistics

32

Suomen kielitieteellisen yhdistyksen aikakauskirja
Tidskrift för den Språkvetenskapliga föreningen i Finland
Journal of the Linguistic Association of Finland

Editors:

Johanna Isosävi
Santeri Palviainen

Antti Kamunen
Tomi Visakko

Aleksi Mäkilähde

Copy editor:

Tommi Nieminen

Advisory editorial board:

Werner Abraham
University of Vienna

Arto Anttila
Stanford University

Kimmo Granqvist
Södertörn University

Auli Hakulinen
University of Helsinki

Jaakko Hämeen-Anttila
University of Edinburgh

Martin Haspelmath
*Max Planck Institute for
Evolutionary Anthropology*

Marja-Liisa Helasvuo
University of Turku

Anders Holmberg
Newcastle University

Tuomas Huumo
University of Turku

Juhani Härmä
University of Helsinki

Fred Karlsson
University of Helsinki

Leena Kolehmainen
University of Turku

Meri Larjavaara
Åbo Akademi University

Jaakko Leino
University of Helsinki

Matti Miestamo
University of Helsinki

Urpo Nikanne
Åbo Akademi University

Martti Nyman
University of Helsinki

Krista Ojutkangas
University of Turku

Mirja Saari
University of Helsinki

Ulla Tuomarla
University of Helsinki

Maria Vilkuna
University of Helsinki

Jussi Ylikoski
*University of Oulu,
Sámi University of
Applied Sciences*

Jan-Ola Östman
University of Helsinki

The editors wish to acknowledge the invaluable work of the advisory editorial board and to express their gratitude to those leaving the board this year, to those continuing to serve in the board as well as to those now joining the board.

2019

ISSN-L: 1456-8438
ISSN: 1456-3438 (Print)
ISSN: 1796-279X (Online)

Painosalama Oy, Turku 2020

Contents

Reviewers of *SKY JoL* 32 (2019) 5

Saleem Abdelhady

📖 A sociolinguistic study of the use of the definite article (*ʔal*)
in the Jordan Valley 7–41

Marta Coll-Florit & Salvador Climent

📖 A new methodology for conceptual metaphor detection and
formulation in corpora: A case study on a mental health corpus 43–74

Maarja-Liisa Pilvik

📖 Assessing the productivity of the Estonian deverbal suffix
-mine in five registers of Estonian 75–103

Hanna Pook

📖 The pronoun *kes* ‘who’ and its referent’s animacy in
Estonian dialects 105–144

Ulla Vanhatalo & Heli Tissari & Taru Lilja &

Kimmo Vehkalahti & Mari Siirainen

📖 “Something bad can now happen to me here”:
Meaning components of emotion words 145–179

Book reviews:

Abdelkader Fassi Fehri. 2018. *Constructing feminine to mean: Gender, number, numeral, and quantifier extensions in Arabic*
Reviewed by Abdel Rahman Mitib Altakhaineh & Aseel Zibin . . . 181–186

Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.). 2017.
Commands: A cross-linguistic typology: Explorations in linguistic typology
Reviewed by Rea Peltola 187–194

Pilar Garcés-Conejos Blitvich & Lucía Fernández-Amaya &
María de la O Hernández-López (eds.). 2019.
Technology mediated service encounters
Reviewed by Heli Tissari 195–201

Reviewers of *SKY JoL* 32 (2019)

Mahmoud Al-Khatib (Jordan University of Science and Technology), Peter Hallman (Austrian Research Institute for Artificial Intelligence), Annekatrin Kaivapalu (University of Eastern Finland), Eino Koponen (University of Oulu), Andreas Musolff (University of East Anglia), Urpo Nikanne (Åbo Akademi University), Edgar Onea (University of Graz), Hannu Remes, Cristina Soriano (University of Geneva), Maria Vilkuna (University of Helsinki), Zhengdao Ye (Australian National University), Islam Youssef (University of South-Eastern Norway)

The editors are most grateful to all the scholars who have acted as reviewers for *SKY Journal of Linguistics* in 2019, including those who wish to remain anonymous and are therefore not listed here.

A sociolinguistic study of the use of the definite article (*ʔal*) in the Jordan Valley

Saleem Abdelhady
Memorial University of Newfoundland

Abstract

This study investigates the variable use of the definite prefix (*ʔal*) in the Arabic spoken in Ghawareneh, a community in the Jordan Valley, Jordan. In Standard Arabic, the definite prefix undergoes assimilation only before coronal sounds; in Ghawareneh Arabic, assimilation is variably possible before all sounds, including non-coronals. The phenomenon is analyzed from a sociolinguistic perspective. Data is collected through informal interviews. The speech sample consists of the naturally occurring speech of sixteen speakers (2 genders × 2 education levels × 2 age groups × 2 speakers per cell), stratified by age, gender, and level of education. The study analyzes the data against lexical, phonological and sociolinguistic factors. The study has shown the impact of phonological factors on the assimilation of glottal consonants. The semantic content of words restricts the use of the non-standard variant. The education, age, and gender factors show that the non-standard variant is more likely to be used among uneducated, old, and male speakers. The study concludes that Ghawareneh speakers are moving linguistically away from the Jordan Valley to achieve some level of urbanization by avoiding the non-standard variant. Both the attitude toward life in the Jordan Valley and the degree of contact with urban centers and with other communities may impact the variable use of the definite prefix. However, for older generations, the use of the non-standard variant preserves and solidifies their identity.

Keywords: language variation and change, assimilation, definite article, Jordanian Arabic, lexicalist hypothesis, coronals

1 Introduction

This study investigates the variable use of the definite prefix (*ʔal*) in the Arabic spoken in Ghawareneh, a community in the Jordan Valley, Jordan. In Standard Arabic (SA), the definite prefix undergoes assimilation only before coronal (“sun”) sounds; in Ghawareneh Arabic (GA), assimilation is variably possible before all sounds, including non-coronals (“moon letters”). This peculiar

linguistic phenomenon is embedded within a complex social structure that incorporates several tribal roots and communities, including Suqour (Bedouin origin), Ghawareneh (rural farming origin), and others. Ghawareneh is used as a general term to cover all people who inhabit the Jordan Valley. However, in its actual use, Ghawareneh, in fact, refers to only one ethnicity, which can be distinguished from *alʔahraar* ‘the freemen’, and *ʔalʔabiid* ‘the slaves’. According to Bani Yasin (1980), the Jordan Valley had been structured based on tribal origins. While such distinction no longer exists, Ghawareneh is still used to refer to dark-skin people (Bani Yasin 1980), living in the area that extends between the Jordanian and Palestinian borders (see § 3.1 for further details). Even though Ghawareneh, Suqour and other communities interact on a daily basis, every community has pride in its origin and dialect. The interaction of Ghawareneh with other communities that follow the SA assimilation rules, and the closeness of the Jordan Valley to urban centers, leave a question on the sociolinguistic status of the definite prefix in that region. In this study, I analyze this phenomenon from a sociolinguistic perspective to highlight any linguistic and social factors that might lead to the use of one variant over the other. The study covers a range of speakers of different age groups.

The paper is organized as follows. The first section presents the background about the status of the definite article in SA and across dialects and introduces the factors that pertain to language variation. The second section highlights the problem of the study and its aims and hypotheses and introduces the framework followed in data collection and analysis. A thorough analysis and discussion are provided in sections 4 and 5. The final section concludes the study by presenting the significant findings and suggesting new arenas for further research.

2 Background

The aim of this section is to provide an understanding of the interaction of the definite article with lexical items. First, I present a brief description of the morphological and phonological status of the definite article in SA. Then, a brief description of the phonological status of the definite article across dialects is presented. The following section introduces a number of social factors that affect variation and connects such variables with studies of variation in the Arab world.

2.1 The definite article in SA

The definite article *ʔal-* is used as a prefix in Arabic to specify the grammatical definiteness of lexical items. For example, nouns such as *kitaab* ‘book’ and adjectives such as *kabiir* ‘big’ can be specified by attaching the definite article to the left of the lexical item to be realized as *ʔal-kitaab* ‘the book’ and *ʔal-kabiir* ‘the big’, respectively.

Arab grammarians have provided a description of how the definite article is pronounced before some letters. Based on their description of the phenomenon, Arabic letters are divided into two groups: sun letters and moon letters. In sun letters, the letter *l* of the definite article assimilates to the sound of the following sun letter; while in moon letters, it does not show any assimilation. The letters are grouped as follows:

1. sun letters: /t/ ت /ʔ/ ط /d/ د /d/ ظ /θ/ ث /ð/ ذ /ð/ ض /s/ س /s/ ص /z/ ز /ʃ/ ش /l/ ل /n/ ن /r/ ر
2. moon letters: /b/ ب /dʒ/ ج /k/ ك /q/ ق /ʔ/ ع /f/ ف /x/ خ /ʁ/ غ /h/ ح /ʕ/ هـ /m/ م /w/ و /j/ ي

Thus, in lexical items such as *fams* ‘sun’, the definite prefix is pronounced as *ʔaf* as in *ʔaf-fams* / **ʔal-fams* ‘the sun’, while in lexical items such as *qamar* ‘moon’, the prefix is pronounced as *ʔal* as in *ʔal-qamar* / **ʔaq-qamar* ‘the moon’.

From a phonological perspective, the sounds of sun and moon letters are distinguished in terms of their coronality (Kenstowicz 1994). In SA and the majority of dialects spoken, coronal sounds (sun letters) induce complete assimilation of the definite prefix, resulting in a geminate coronal consonant while non-coronal ones (moon letters) do not (Kenstowicz 1994; Ryding 2005; Heselwood & Watson 2015). From an acoustic point of view, Heselwood & Watson (2013; 2015) argue against the view that calls for looking at the interaction of the lateral /l/ of the definite article with coronal sounds as a synchronic assimilation process since the process does not fulfill one criterion of synchronic assimilation: optionality – the process is not optional as when /l/ occurs within or across a word boundary, as in *ħabil rafiiʕ* ‘a thin robe’ in Syrian Arabic (Heselwood et al. 2011; Heselwood & Watson 2013: 34). The study provides a new perspective for the interaction of the lateral /l/ of the definite article with coronal sounds. Nevertheless, whether the process is synchronic assimilation or not is beyond the scope of this paper.

2.2 The definite article across dialects

The phonology of the definite article and its assimilatory process has been an area of investigation in the literature (Bani Yasin 1980; Watson 2002; Assiri 2008; Al-Qenaie 2011; Heselwood & Watson 2013; 2015). In the dialects, the article can vary in three different ways: whether the article begins with a glottal stop /ʔ/ or not, whether the vowel of the definite article is /a/ or /i/ and whether the lateral /l/ undergoes assimilation or behaves differently.

While the lateral /l/ of the definite article assimilates to coronal consonants in SA, it does not always follow this pattern across dialects (Bani Yasin 1980; Assiri 2008; Heselwood & Watson 2013; 2015). Two different patterns can appear: ignoring the phonological environment by replacing /l/ with a fixed segment once it interacts with coronal and non-coronal sounds and/or overapplying the assimilation rule – *l* assimilates to the following sound regardless of whether the sound is a coronal or not.

In the Ghawareneh community (a community living in the Jordan Valley, Bani Yasin 1980) and some dialects that are found in the Western Yemeni mountains and Southern Oman (Behnstedt 1987: 85, cited in Heselwood & Watson 2013: 18), the definite article assimilates to any following consonant (Bani Yasin 1980: 217). Assimilation of the definite article to non-coronal consonants is also attested in some varieties of Moroccan Arabic (Heath 2002) and in Christian Baghdadi Arabic (Abu-Haidar 1991). According to Bani Yasin (1980) and Behnstedt (1987: 85), cited in Heselwood & Watson (2013; 2015), the process of overapplying the assimilation rule results in “an article that involves gemination of any nominal-initial consonant”.

The second pattern that appears in some dialects is the substitution of the lateral /l/ of the definite prefix with a fixed segment regardless of the initial consonant of the lexical item. In Yemeni Majz (spoken in Yemen), the lateral /l/ is substituted by /n/ regardless of its phonological environment, as in *ʔin-ša‘bah* ‘the female donkey foal’ and *ʔin-fams* ‘the sun’ (Behnstedt 1987: 85, cited in Heselwood & Watson 2013: 35). Another substitution process involves replacing the lateral /l/ with /m/. In Rijāl Alma’ (a dialect spoken in Saudi Arabia), the definite article does not assimilate to any sound (Assiri 2008). Its lateral /l/ is substituted with /m/ regardless of the following environment, as in *am-safar* ‘the journey’ and *am-qamar* ‘the moon’.

Thus, we see three possible assimilation processes: the lateral /l/ of the definite article can assimilate to a coronal; it can be substituted by a fixed segment regardless of the phonological environment, or it can assimilate to

any phonological environment.

In addition, across dialects, the vowel of the definite prefix varies. In some dialects, a high front vowel is used instead of the low front one. In Kuwaiti Arabic, *ʔal* is realized as *ʔil* or *il* (Al-Qenaie 2011). In this dialect, it is not only vowel raising that differs from SA but also deletion of the glottal stop. For example, lexical items such as *ʔintixabaat* ‘election’ is realized as *lintixabaat* ‘the election’. According to Al-Qenaie (2011: 241), the phonological structure of the lexical item passes through the following stages before the output is thus realized. In the underlying form: first, the lexical item is prefixed to *ʔal-intixabaat*, then the vowel is raised, as in *ʔil-intixaba*. Then, under the process of aphaeresis, the glottal stop and the vowel are deleted: *l-intixabaat*. Finally, the word undergoes re-syllabification to be realized as *lintixabaat*. The phonological decomposition under the impact of the definite prefix is argued to depend on linguistic and social factors that are discussed in Section 3 in more detail.

Between SA and spoken Arabic dialects, the definite prefix has six allomorphs as argued by Haywood & Nahmad (1965: 22), cited in Heselwood & Watson (2015: 158); with non-coronal consonants or vowel /ʔal/, /ʔil/ or /l/ is used. With coronal consonants, the lateral l becomes similar to the following coronal consonant. Thus, three patterns appear /ʔaC-C[cor]/, /ʔiC-C[cor]/, and /C[cor]/. The six patterns are argued to be governed by linguistic and social contexts (Heselwood & Watson 2015: 158).

2.3 Lexicalist and phonological variables

Based on studies on language variation and change, linguists noticed that phonological variation could be restricted to certain lexical items and not others (Abdel-Jawad & Suleiman 1990). This has been referred to in the literature as the lexicalist hypothesis. The lexicalist hypothesis states that some speakers may avoid using non-standard variants with certain lexical items, they may associate the non-standard variant with other lexical items, and they might alternate between the standard and non-standard forms for other lexical items.

In Jordanian Arabic, Abdel-Jawad & Suleiman (1990: 298) divide lexical items into three categories. The first class of lexical items includes technical, educated and cultivated words. Based on their study, they claimed that in such a class of words, the standard variant would more likely be used. They exemplify this class of words by lexical items such as *ʔiqtiṣaad*

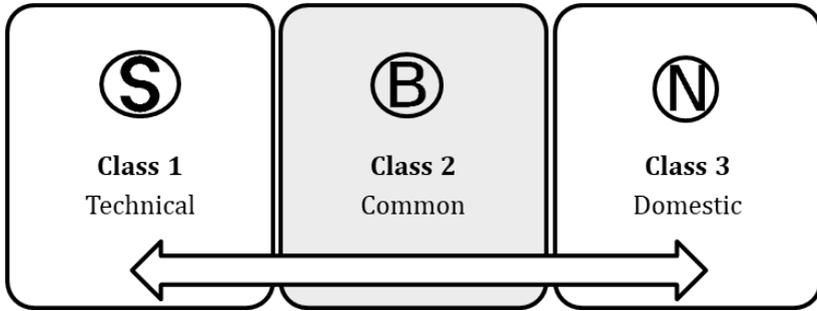


Figure 1. Illustration of the semantic continuum¹

‘economy’, *maqsam* ‘switch’ and the like. The second class of lexical items includes “etymologically related, basic, shared common items which may be morpho-semantically, and phonologically (a) identical or (b) not identical” (ibid.). This class of words is argued to use both the standard and non-standard variants in a random way. Lexical items such as *qamar* ‘moon’ and *ḍahab* ‘gold’ are representative of the former. Lexical items such as /kaif/ ‘how’ (in SA) >> /kiif/ ‘how’) illustrate the latter. The third class of lexical items consists of words that refer to “domestic and local objects and concepts” such as pure dialectal items that do not have standard synonymous words (e. g., *ṭṣabbara* ‘fire place’), blended lexical items *haaḍa al-waqt* ‘this time’ >> *halqet* ‘this time’ and words that are synonymous with standard words such as *bard* ‘cold’ (standard) >> *sagaḥa* ‘cold’ (colloquial). In this class of words, the non-standard variant is more likely to appear. The use of certain variants could be conditioned and restricted to certain lexical items. The continuum in Fig. 1 is illustrative.

The lexicalist hypothesis and the phonological contexts are attested in Arabic across a number of dialects and variables. Yet, such hypotheses are not yet examined on the use of the definite prefix *ʔal* ‘the’. Thus, in this study, it would be significant to test the validity of such a hypothesis.

¹ S: standard variant, B: both variants and N: non-standard variant.

2.4 Studies of variation in Arabic

Before dealing with the definite article as a sociolinguistic variable, I review the effect of sociolinguistic and linguistic factors in Arabic speaking communities: the variable use of the definite prefix (Assiri 2008; for further details on current research on variation in the Arabic-speaking world, see Horesh & Cotter 2016). Then, I review the sociolinguistic status of variation in Jordanian Arabic (e. g. Abdel-Jawad 1981; Al-Wer 1991; Al-Tamimi 2001; Al-Shatarat 2015; Abu Ain 2016; Omari & Van Herk 2016).

In Saudi Arabic, Assiri (2008) investigates the impact of age, gender and education on the use of two phonological variables: the variable alternation between *-il* (standard variant) – *-im* (non-standard variant) and [k] (standard variant) – [x] (non-standard variant) in Rjaal Almaṣ, Saudi Arabia. He finds that sociolinguistic factors affect the alternation between the use of standard and non-standard forms, yet the two variables are not alike (Assiri 2008: 42–45). With regard to the use of the definite article, he finds that the non-standard variant of the definite article, *-im*, is used more by young males, while educated speakers, regardless of their gender and age, tend to avoid that variant. Even though Assiri's (2008) study is the first study to highlight the variable use of the definite prefix, it is not clear whether the choice of non-standard variants is affected by the semantic content of lexical items or not, nor does he provide a reason to account for the use of the nasal sound, [m] over other sounds.

In Jordanian Arabic, studies on language variation and change have been abundant and have tackled different domains: urban centers (Abdel-Jawad 1981; Al-Khatib 1988; Al-Wer 1991), Palestinian refugee camps (Al-Shatarat 2015), rural areas (Abu Ain 2016), and the speech of rural immigrants to urban centers (Al-Tamimi 2001). Studies on Jordanian Arabic show that females are the ones who use urban variants, educated speakers use standard variants more than other speakers, and that young speakers are usually the ones who lead a change (cf. Abdel-Jawad 1981 and Al-Khatib 1988). Nevertheless, variation in the Jordan Valley has been overlooked. In fact, as mentioned earlier (see section 1), only one study has explored the dialects spoken there (Bani Yasin 1980), and that study dealt with the dialects spoken from phonological, semantic, and syntactic perspectives only. Therefore, it is not clear up to this date what linguistic changes took place from 1980 under the effect of social and linguistic factors.



Map 1. The location of the Jordan Valley. Adapted from: “Location Map of Jordan”, by NordNordWest, https://commons.wikimedia.org/wiki/File:Jordan_location_map.svg (accessed 2019-03-08). Licensed under CC BY-SA 3.0 DE (<https://creativecommons.org/licenses/by-sa/3.0/de/deed.en>). Image changed to greyscale. Labels added for clarification.

3 Methodology

First, I present the setting of the study with some geographical, demographic, and political information. Presenting such information helps in forming an idea about the structure of communities living there. Then, I highlight the problem of the study, its hypotheses, and its aims. After that, the population and the sample of the study and the rationale behind choosing such a sample are stated. Finally, I introduce the framework followed in data collection and analysis.

3.1 Regional setting and ethnographic background

The Jordan Valley (Al-Ghor) is a valley located between Palestine to the west and Jordan to the east (see Map 1). It extends for almost 105 km (the Jordanian part begins just 30 km from Irbid city to the west, passes through the Dead Sea and reaches up to Aqaba city in the south).



Map 2. The location of Mashari. Adapted from: “Proposed September 2019 Israeli annexation of Jordan Valley”, by Nice4What (adapted from a map by NordNordWest), https://commons.wikimedia.org/wiki/File:Proposed_September_2019_Israeli_annexation_of_Jordan_Valley.svg (accessed 2019-03-08). Licensed under CC BY-SA 3.0 (<https://creativecommons.org/licenses/by-sa/3.0/deed.en>). Image changed to greyscale. A circle has been added to indicate where the study took place.

The Jordan Valley is the lowest area in the region, with a year-round warm climate that is warmer than the rest of the country. Such a climate makes farming a prominent job in the area.

The valley is divided into many sub-regions, from north to west: North Shuna, Mughair, Waqas, Shaykh Husayn, Al-Mashari, Wadi Al-Ryan, Kuraymeh, Sawalha and others. The hollow circle shows where the study took place (see Map 2).

The Jordan Valley is inhabited by a wide array of communities. All communities living in that area, nevertheless, are referred to as Ghawarneh by Jordanian People. According to Bani Yasin (1980), in its broader sense, Ghawarneh refers to three ethnic groups: *alʔahraar* ‘the freemen’, the Ghawarneh, and *ʔalʔabiid* ‘the slaves’. These groups represent the original inhabitants of the Jordan Valley. Bani Yasin (1980: 29) provides a detailed

description of Ghawarneh.

All the people of Al-Ghor are sedentary. They no longer live in tents, except temporarily near their own land at the time of active cultivation. City dwellers call them *fallahin* or *Ghawarneh* and think as [sic] them unused to modern ways of life, like entertainments such as the cinema. In fact, the lifestyle of the people of Al-Ghor is very traditional. Thus, the [sic] women of Al-Ghor do not go out of their houses without covering their heads, according to Islamic Law, but they go out and work in the fields alongside the men. As far as social structure is concerned, their life is more or less like the life of Bedouins. Each tribe in Al-Ghor has a chief or sheikh, who is responsible for everything.

The description reflects the social status of Ghawarneh and the way they are looked at by other communities who live in cities. Most of Ghawarneh belong to the low working class. They prefer the rural style of living. Moreover, the description shows some aspects of their social ties. People in that community have strong blood/tribal ties. Their social structure imposes certain expectations on how individuals should behave or speak in that community. The description extends and captures how other communities look at Ghawarneh. It is enough for “city dwellers” to look down upon people if they are living in the Jordan Valley, regardless of their level of education, occupation, or economic status. If a person lives in the Jordan Valley, he is taken to be uncivilized.

3.2 Problem, hypotheses, and limitations of the study

Bani Yasin (1980) draws attention to the phonological status of the definite article in the speech of the Ghawarneh community. Nevertheless, he shows that the assimilation process is not very well understood across the community, and he does not provide a reasonable track of how this process started. Thus, it is not clear if the process has undergone any change with its interaction with other dialects and communities. Bani Yasin (1980: 220) concludes that

Al-Ghor dialect has its own characteristics with regard to phonetic realization of the definite article, *in which matter it is quite unusual* (italics mine). Any traces of more specific derivations from either the Ghaza region, or parts of Saudia [sic] Arabia, as maintained in tribal traditions are no longer observable, having been merged with other dominant linguistic traditions or developments, within Al-Ghor.

His observation highlights the fact that speakers of this community are more likely to affect or be affected by “other dominant linguistic traditions” (1980: 220). Even though the problem is stated, it has been overlooked in the literature.

It is hypothesized that (1) the non-standard form is favored by speakers who match the demographic profile of non-standard speakers in other studies (e. g. Abdel-Jawad 1981; Assiri 2008 and others), (2) the non-standard variant will target all non-coronal sounds, and (3) the semantic content and frequency of words can affect variation.

Covering all aspects and testing the effects of all social factors would require more space and time. Thus, the study is limited in many respects:

1. It is likely that political and historical factors affect variation in this community. However, these factors, among others such as ethnicity, social status, and style/register and the effect of phonological factors such as pauses, emphasis, stress, hesitation, and syntactic position of lexical items and the like, are left for further studies.
2. The sample of the study covers ages between 19 and 87. It would be a point of investigation to see how children acquire/use such a variable.
3. The study is limited to the Ghawarneh community and does not cover all sub-regions of the Jordan Valley. It would also be worthwhile to widen the scope to cover the southern part of the Jordan Valley.
4. The study is limited to non-coronal sounds attested in the literature, excluding borrowed sounds like [v] in lexical items, such as *viideo* ‘video’.

3.3 Population and sampling

The population of the study represents the speech of the Ghawarneh community living in the Jordan Valley. Working as a schoolteacher with the United Nations Refugees Working Agency (UNRWA) in Al-Mashari and Kuraymeh primary schools and mixing with the Ghawarneh community for over a year in 2009 makes me an in-group member to some degree. Thus, the sample of the study is chosen based on my own social network. The speech sample consists of the naturally occurring speech of sixteen speakers (2 genders × 2 education levels × 2 age groups × 2 speakers per cell), stratified

by age, gender, and level of education. Age includes old (50 and above) and young categories (below 50). Gender is categorized based on biological sex: males and females. The level of education is based on whether the speaker has received any post-secondary degree or not. Speakers who have received primary schooling are classified as uneducated. The reason for this classification is due to the structure of the educational system in Jordan. The Ministry of Education makes basic schooling (pre-primary, primary, and secondary levels) available in rural areas. Higher education is only available in urban centers.

3.4 Data collection and analysis

3.4.1 Data collection

Data was collected through informal sociolinguistic interviews, recorded using a Samsung Note 4 mobile phone. The interviews were conducted in different settings, including homes of participants and coffee shops.² The interviews are of an average of 30 to 35 minutes each. Most interviews are conducted by me with the speakers' family members or friends present following Memorial University ethics procedures.

The strategy followed in conducting most interviews was almost the same for each participant:³ I asked the same questions, not in exact order, and emphasized areas in which I felt that the participant was able to talk spontaneously. The technique followed in this case has been controlled not only by the questions asked but also by the conversational frames of schematic discourse. Such frames aimed to activate topics that are related to common ground knowledge and experiences shared with me. For example, while talking about the fasting month of Ramadan, a speaker would mention the name of one of the well-known TV series. By using the conversational framing technique, I maintained the flow of conversation by asking further details

² Even though *diglossia* exists in Arabic communities, the use of informal interviews in very informal settings targets the low variety of spoken Jordanian Arabic (the vernacular). The high variety, the variety that is used in certain specific settings such as Muslim Friday sermon or formal education but not ordinary conversations (Ferguson 1959), is excluded. The term *standard*, therefore, refers to the way the definite prefix is used in cities and most modern vernacular Arabic varieties (see section 2.1).

³ The recordings of participants who have been able to talk naturally about topics they felt competent were shorter with less questions asked, yet they fulfilled the purpose of the conversation.

about the story of this series and his/her impression. This technique adds to the naturalness of conversations.

The interviews were about multiple topics: farming, dreams, childhood, the fasting month of Ramadan, education if any, marriage, traditional dishes, and the way they are made, life in the Jordan Valley as compared to that of a city, such as Irbid and Amman. The interview guiding questions are about 60 questions (see Appendix A). The choice of such topics and the presence of family members helped in eliminating the *observer's paradox* (Labov 1972).

The aim of choosing such topics is related to the nature of life in the Jordan Valley and the fasting month of Ramadan; speakers can talk naturally as the topics are accessible to all. Moreover, in Arabic, the topics would naturally target non-coronal sounds. For example, farming would spontaneously trigger lexical items that are related to different types of fruits and vegetables that are grown in that area, such as *burtugaal* 'oranges', *fiul* 'beans', *hamḍijaat* 'citrus fruit', *muuz* 'bananas' and others. Talking about life in the city versus that in the Jordan Valley triggers lexical items such as *ʔor* 'the Jordan Valley', *ḥajaah* 'life', *maʔaariʔ* 'Mashari', *qanaah* 'the canal' and the like. Talking about the fasting month of Ramadan would trigger words that are associated with prayers such as *mayrib* 'evening prayer', *ʔaʔir* 'afternoon prayer', *ʔiʔa* 'night prayer', *ʔadʔir* 'dawn prayer', *ʔaʔuur* 'meal that breaks fasting', *qurʔaan* 'The Holy Quran', *bab al-ḥaarah* 'Bab Al-harah, a popular series in Ramadan' and the like. The use of these words in the course of speech would trigger a natural use of the definite prefix with non-coronal sounds.

Albeit part of Muslims' common ground, some topics that are related to religious traditions are difficult for some. Thus, in line with Memorial University ethics procedures, participants were given the freedom to change the topic of the conversation and/or the question asked, to pause the recording at any time, or to stop it if deemed necessary.

3.4.2 Data analysis

After conducting an interview, I replayed the conversations multiple times to make sure that I extracted every token of the variable (article + non-coronal sound) used by participants. Tokens were coded for the variant used, speaker demographics (age, gender, and level of education), and lexical-semantic content (Technical "T", Common "C" and Dialectical "D"), based on my intuition as a speaker of Jordanian Arabic and the views of my Ghawarneh friends.

Goldvarb X (Sankoff et al. 2005) is used for statistical analysis. Binomial, Up, and Down test is run to obtain the degree of significance of each factor on the realization of variables across all non-coronal sounds. Cross-tabulation is also used to obtain the relationship between different intersecting factors.

4 Findings and Discussion

I begin by presenting the results in three stages based on the questions presented earlier: (i) which of the non-coronal sounds are subject to change, (b) what is the impact of the semantic content on LVC and (c) what is the impact of demographic factors on change? Then I proceed to discuss and attempt to explain the phenomenon.

The analysis of data shows the use of the definite article in 1820 lexical items – with the exclusion of words in which the definite article is not clearly recognized. The assimilation to non-coronal sounds shows that the non-standard variant appears in 642 (35.3%) tokens while the standard one is used in 1178 (64.7%) tokens. To understand the alternation of using one variant over the other, the data is further analyzed under the impact of linguistic and non-linguistic factors.

4.1 The impact of linguistic factors

Considering the assimilation and non-assimilation of individual non-coronal sounds, the analysis shows the distribution in Table 1. The table shows that the non-standard variant is seldom used with lexical items that begin with glottal stops, as [ʔ] and is less likely to appear in those that have glottal fricatives, as [h]. Albeit used often in most lexical items, the standard variant is less used with the affricate consonant, [dʒ].

A closer look at the distribution of sounds based on their voicing features (Table 2) shows the significance of such features in the variable use of *ʔal*. The lateral [l] of the definite prefix assimilates to voiced non-coronal sounds at a factor weight of .57 more than voiceless ones, .39.

Manner features (Table 3) show almost an identical relative strength of .31. As can be seen from the table below, the most frequent category that undergoes assimilation is affricates, while the least frequent one is stops. Apart from the affricate-coronal consonant, the relative strength (range) of manner features drops down to almost .11.

Table 1. Factors favoring assimilation of the definite article in Ghawareneh Arabic: the place of the initial consonant of the following word

Sound		% non-std	N
dʒ	.91	59.1	93
b	.64	44.8	212
m	.63	37.8	423
ʕ	.61	43.5	207
k	.56	40	55
f	.52	35.1	94
x	.51	37.8	45
ħ	.49	31.6	206
w	.48	33	97
q	.45	36.5	85
ɣ	.41	35.2	71
j	.27	34.3	35
h	.18	11.1	18
ʔ	.04	1.7	179
			1820

Corrected mean .227, range 87

Table 2. Factors favouring assimilation of the definite article in Ghawareneh Arabic: the voicing of the initial consonant of the following word

Voicing		% non-std	N
Voiced	.57	41.6	1138
Voiceless	.39	25.4	682
			1820

Corrected mean .39, range 18

Table 3. Factors favouring assimilation of the definite article in Ghawareneh Arabic: manner features of the initial consonant of the following word

manner features		% non-std	N
affricates	.72	59.1	93
nasals	.53	37.8	423
fricatives	.51	36.2	641
glides	.48	33.3	132
stops	.42	28.4	531
	.39		1820

Corrected mean .39, range 31

Table 4. Factors favouring assimilation of the definite article in Ghawareneh Arabic: place features of the initial consonant of the following word

place features		% non-std	N
coronal	.73	59.1	93
labial	.55	39.5	729
dorsal	.51	35.8	388
laryngeal	.40	26.2	610
			1820

Corrected mean .39, range 33

Table 5. Factors favouring assimilation of the definite article in Ghawareneh Arabic: lexical class of the following word

lexical item		% non-std	N
dialectal	.71	57.3	405
common	.49	32.3	1219
technical	.19	8.2	196
			1820

Corrected mean .39, range 52

As non-coronal sounds are hard to be grouped into natural classes, the statistical analysis of place features shows almost a similar outcome with manner features. See Table 4. Place features (.33) show almost an identical relative strength with the manner features (.31). The most varied category to undergo assimilation is coronals; similar to most non-coronal consonants, the affricate-coronal consonant [dʒ] is singled out by some place and manner categories of its own. In fact, the significance of manner and place features seems to be in part due to the strong effect of the affricate-coronal consonant [dʒ] that has a factor weight of .72 in each category. The second category to assimilate is labials. Laryngeals assimilate the least.

The semantic content of words is found to impact the alternation between standard and non-standard way of assimilation. Table 5 illustrates the distribution of both variants across three lexical classes based on the lexicalist hypothesis.

Table 5 shows that the non-standard variant is rarely used with technical words. Yet, its distribution between common and dialectical lexical items is evident, and it is more likely to appear in the latter.

It should be noted that while variation could be accounted for in terms of semantic meaning, the use of technical words is noticed to overlap with social factors such as education.

Bani Yasin (1980), among others, has noticed that the definite article in the Jordan Valley assimilates to non-coronal sounds. Such an observation, however, is too general and has not provided any clue on which of non-coronal sounds is used/not used often. The data show that not all non-coronal sounds affect the use of one variant over the other in the same degree. In fact, some sounds are almost used exclusively with the standard variant. Thus, there should be some relation between the sound used and the type of variant.

The use of the standard variant with [ʔ] can be attributed to the fact that using the non-standard form would result in three glottal stops in the lexical item. For example, in lexical items such as *ʔamiir* ‘prince’, the assimilation to non-coronal sounds would result in triple glottal stops, *ʔaʔ.ʔamiir* ‘the prince’ which makes an unnatural sound combination. Thus, I argue that if the process makes the pronunciation difficult, speakers refrain from using the non-standard variant. The same argument can be extended to the glottal fricative [h]. It seems that since [ʔ] is used at the beginning of the definite prefix, it bans the use of glottal consonants.

A natural conclusion to put forward is that under the impact of the Obligatory Contour Principle, a principle that bans “consecutive identical

features” (Odden 1986) or “nearby segments that are similar or identical from occurring” across morpheme boundaries (McCarthy & Prince 1995: 92), the use of the non-standard variant would be restricted. Such an argument has been observed to affect some instances of reduplication in some languages such as Akan, a language spoken in Ghana, in which a sequence of at least three coronal segments blocks the phonological environment of reduplication; two segments, nevertheless, are allowed (McCarthy & Prince 1995: 92). The OCP principle, I argue, would not be active to all non-coronal sounds as the assimilation to such sounds would not result in three consecutive identical segments/features, which is an exceptional application of the OCP principle.⁴ The few instances observed are perceived to behave more like deletion of the lateral [l] of the definite prefix, rather than a full gemination of glottal stops.

In addition, based on the inventory of non-coronal sounds (Heselwood & Watson 2013: 35) and studies on Arabic grammar (Aala Addin 2016) the affricate coronal sound [dʒ] and its allophonic variant [ʒ] is classified as a non-coronal sound that does not undergo assimilation.⁵ Therefore, while a coronal consonant should be treated as a sun letter due to its place of articulation, the sound does not belong to sun letters. What the actual realization and the status of this sound in the speech of Ghawarena shows is that the sound is not clearly defined as a non-coronal sound in this dialect: almost 60% of tokens have undergone assimilation, a higher rate than for other sounds. The rule of assimilation, therefore, seems to be acting in accordance to the natural way of assimilating the lateral [l] to coronals; the sound fits those original specifications of coronal sounds; it is pronounced with the tip of the tongue in a similar way that the coronal sound [d] is pronounced.

However, this sound does not categorically assimilate, in the way that regular coronals do. I argue that this is because this sound at the very first levels of education is instructed to be pronounced as a non-coronal sound, in which the definite prefix does not assimilate. The non-assimilation, therefore, would be attributed to language maintenance in which this sound has to be set and pronounced as a non-coronal sound. The sound highlights a significant

⁴ Partial assimilation could invoke less violation of OCP. It is worth exploring if partial assimilation of the lateral [l] (devoiced [l̥]) and glottalized/pharyngealized [l̥ʕ]) occurs from an acoustic point of view.

⁵ From a historical-phonological perspective, the sound /dʒ/ has sometimes been treated as a non-coronal sound [g]. Some traces of the non-coronal use of this sound can be found in varieties such as Egyptian Arabic (EA) and Omani Arabic. For example, in EA, words that take [dʒ] in SA are pronounced with [g] instead, cf., *ʔal.dʒunuud* ‘the soldiers’ (SA) vs. *ʔal.gunuud* ‘the soldiers’ (EA). See Woidich & Zack (2009).

bottom-line for a correlation between Classical Arabic and Ghawareneh Arabic (and other varieties of Arabic, such as Omani Arabic and Yemani Arabic); while it is unquestionable that /dʒ/ is a coronal sound, the data show instances of a covert contrast between a coronal and a non-coronal sound,⁶ a voiced palatal plosive /j/. At the perception level, the sounds are hard to set apart. Thus, the non-assimilatory patterns of /dʒ/ might be manifestations of a change in progress from below, that is, in the speech of Ghawareneh, the two sounds might be present at their phonetic inventory. Because of the fact that these sounds are perceptually identical, all treatments of the non-coronal sound, represented by the letter <ج>, capture only one variant of this sound /dʒ/ and overlook the other covert contrast with /j/. The data provide a piece of evidence toward the presence of /j/ which might be manifested through the cases in which the definite prefix is not assimilating to /dʒ/. Since the contrast is covert, I argue that the language is undergoing a change in progress from below the level of awareness of Ghawareneh speakers; that is, Ghawareneh are going from the marked /j/ to the unmarked /dʒ/. Because of the change toward the unmarked, the impact of the rule of assimilation is evident.⁷

The use of the non-standard variant with the rest of the sounds ranges between 30% and 45%. As these sounds do not make a natural class by themselves, it would be hard to predict what feature accounts directly for their use. However, based on grouping sounds to their prominent voicing and points of articulation and manner features, the study shows that voiced non-coronal sounds are more likely to trigger assimilation. The reason, I argue, is attributed to markedness (Hayes & Steriade 2004; Rice 2007; Zhang & Tian 2015); even though the exact definition of markedness is debatable in the literature and subject to cross-linguistic variation (see Rice 2007), according to some phonologists marked sounds can be triggers (more than targets) of assimilation, they are unlikely to be epenthetic and subject to neutralization (see Rice 2007: 80 for more defining features). Given the debate over featural markedness, I adapt Lombardi's (1991) argument that calls for defining voiced sounds as

⁶ As pointed by one of the reviewers, *covert contrast* is a concept used primarily in L1 and L2 acquisition of phonology/sounds. Perceptually, a child might not be able to distinguish sounds. The same idea may add to our understanding of instances of *a change from below* in sociolinguistics; perceptually similar sounds might affect the choice of one variant over the other.

⁷ Acoustic analyses will be needed to confirm the distinctions. I recommend further analyses to re-examine the phonetic inventory of speech sounds within spoken Arabic varieties. As the researcher can intuitively tell, the inventory should be expanded since a covert contrast exists between /dʒ/ and /j/. The contrast is highlighted to exist based on the current data, which explains its classification as a non-coronal sound, yet my predictions await to be confirmed.

marked in comparison with their voiceless counterparts. If that is the case, the markedness of the non-coronal sounds with respect to their voicing can affect the tendency and the frequency of using one variant over the other. It should be noted that place and manner features play a role, yet their role is more evident once sounds are taken individually, as discussed above.

It has been noted by Assiri (2008) that some dialects choose to assimilate the definite article to the nasal consonant [m]. He left a question unsettled. Why do speakers choose the non-coronal nasal [m] over other non-coronal sounds? The results of the study show that the most frequent sound to occur, at least in the interviews conducted, in both variants across the sample, is the nasal [m], with 23.2%. This might give a prediction toward why [m] would overapply in replacement of other sounds in some cases such as that of Assiri (2008).

The semantic content of words affects the choice of variants. For example, in words that are related to farming, names of local areas, and objects such as *bayaarah* ‘farm’, *muuz* ‘banana’ and *mafaariʕ* ‘Mashari’ and the like, it is noticed that speakers would more likely use the nonstandard variant. In words that are used in their Classical forms such as *qurʔaan* ‘Holy Quran’, speakers tend to use the standard variant. In addition, in words that are commonly used and do not belong to either of the aforementioned classes, speakers tend to alternate between the two variants. This matches findings of other variables that are used in Irbid City in the North part of Jordan and other studies that deal with LVC from a lexicalist perspective (Abdel-Jawad & Suleiman 1990).

4.2 The impact of social factors

4.2.1 The gender factor

The statistical analysis has shown that gender is statistically significant in the variable use of (*ʔal*). Males tend to assimilate the lateral [l] of the definite article before non-coronal sounds more than females (Table 6).

The results show that female speakers use the standard variant more than male speakers, a pattern that should be “connected with the overall picture of societal structures” (Wodak & Benke 1998) to form an idea toward the underlying reasons that calls women to use the standard variant more than men. Using the standard pattern of assimilation more often by females shows that they are aware that the non-standard variant is interpreted as a form that is used by rural Ghawarneh people; women prefer using the variant that is used

Table 6. Factors favouring assimilation of the definite article in Ghawareneh: gender

gender		% non-std	N
males	.56	40.6	982
females	.43	29.0	838
			1820

Corrected mean .227, range 13

in not only SA but also in urban centers. According to Eckert (1997b) and Trudgill (1972), the use of the standard variant by women is related to some social status or job opportunities that women try to achieve through language. Male speakers, on the other hand, have access to far more ways of defining their status. Thus, using the variant that is more associated with rural aspects of life helps in solidifying the idea that being from a rural agricultural region is part of being hard-working men who are viewed as tougher than those men who work in offices (Trudgill 1972; Assiri 2008).

Labov (1990: 210) states this conclusion as follows:

[The] basic finding can be formulated in two complementary ways: men use more nonstandard forms, less influenced by the social stigma directed against them; [...] women use more standard forms, responding to the overt prestige associated with them.

Even though the gender factor by itself shows a statistical difference, it does not show a full picture of the social structure. Once gender intersects with age, the results show that not all men are behaving differently from women. Young female and male speakers seem to show almost no difference in the use of variants, and both show preference toward the use of the standard variant. Such finding is discussed in the following subsection

4.2.2 The age factor

The age factor plays a significant role in LVC (e. g. Eckert 1997a; Assiri 2008; Habib 2010; Abu Ain 2016). Likewise, in this study, there is a significant statistical difference between young and old speakers: young speakers are more inclined toward using the standard variant while old speakers use the non-standard variant more (Table 7). A cross-analysis of gender and age

Table 7. Factors favouring assimilation of the definite article in Ghawareneh Arabic: age

age		% non-std	N
old	.71	46.5	908
young	.29	22.5	912
	.227		1820

Corrected mean .227, range 42

Table 8. Factors favouring assimilation of the definite article in Ghawareneh Arabic: intersection between age and gender

age	gender	% non-std	N
old	males	54	299
	females	39	137
young	males	23	100
	females	22	106
			1820

(Table 8) shows that old male speakers tend to use the non-standard variant more than old female speakers, but there is no difference in gender between young male and female speakers.

Chambers (2003: 159) notes that “young adulthood is seen as representing a crucial life stage during which standardization increases”. This difference in the variable use of the definite prefix (*ʔal*) among young speakers can be attributed to the following observation. Throughout the interviews conducted, when young speakers were asked about life-style differences between the Jordan Valley and some urban centers such as Irbid city, they indicated that the Jordan Valley is less developed. Thus, they indicated that establishing their lives in a city will provide them with more opportunities for jobs. Thus, the positive evaluation of life in urban centers and the fact that the non-standard variant ties them to the Jordan Valley help in understanding the reason behind choosing the standard variant by young speakers. This conclusion has been supported in many studies (cf. Labov 1972; Habib 2010):

The adult is seen as participating in the standard linguistic market within

the working life stage. Hence, the use of prestige forms is thought to peak in the middle years when the maximum societal pressure to conform is thought to be felt. (Holmes 1992: 186)

Thus, the lack of jobs in the area (the only jobs available in the Jordan Valley are agricultural-based) makes young people interested in working in malls (e.g. Irbid Mall, Arabella Mall, and others) and factories (e.g. Al-Hassan Industrial Estate) found in urban centers such as Irbid, and Amman.

It has been clear that there is almost no difference between young male and female speakers in the variable use of the definite prefix (*ʔal*). The intersection with gender supports the idea that young speakers, regardless of their gender, use the variant that moves them linguistically away from the Jordan Valley. Neutralizing gender shows that young speakers in such a community avoid using the variant that reflects their place of origin. Another reason that might underlie the use of the standard variant could be the conversations conducted. It could be that despite my best efforts to create a comfortable situation and my semi-community status, young speakers might have found themselves in a context for using language from the more formal end of their stylistic repertoire.

For older speakers, the situation has been different. The non-standard variant is found often in the speech of old speakers, with old males using the non-standard variant more than old females. As indicated earlier (see Section 1), the social structure of the Ghawarneh community is complex. It is composed of communities with tribal origins. In the Jordanian society in general and the Jordan Valley, in particular, every community strives to preserve its tribal origin. Under the impact of the vast developments in the Jordanian society and the number of forced migration movements that took place, I argue that the only way to preserve identity and origin is through linguistic means. While the tribal roots are diminishing in younger generations, such roots are still found in the speech of older generations who resist changes.

Three factors inhibit the change toward the standard variant in the speech of old speakers. First, using the non-standard variant would set the Ghawarneh community apart from other communities that use the standard variant; it is, therefore, used a marker to maintain identity. Second, old speakers are more conservative; old informants value family relations and spend most of their time in family gatherings. Thus, the use of the non-standard variant fortifies their identity. Third, it could be that older speakers lacked access to education and interaction with standard speakers and are thus the last generation to

adequately maintain their dialectal forms. Such conclusions are in correlation with other studies on LVC (see Eckert 1997a).

4.2.3 The education factor

The analysis of the education factor shows that uneducated speakers are far more likely to use the non-standard variant (Table 9). Intersecting the results with age (Table 10) shows that uneducated old speakers use the non-standard variable more in their speech, while educated speakers, regardless of age, use the standard variable.

Furthermore, an intersection with gender (Table 11) shows that uneducated male speakers use the non-standard variant more than educated ones. Educated speakers, regardless of gender, show a high tendency toward the use of the standard variant. The following table is representative.

It is crucial to highlight the relation between the education factor and the choice of lexical items for the simple reason that it is assumed that the level of education could affect the choice of lexical items. See Table 12.

Table 12 shows that educated speakers are more likely to use technical lexical items (149 tokens), and only 5% of such words have been used with the non-standard variant. Uneducated speakers, on the other hand, used 47 technical lexical items, 19% of such items assimilate to non-coronal sounds. It is evident that the level of education affects the choice of technical items. Moreover, the low percentages, in both educated and uneducated, show that the semantic factor hinders the assimilation process in technical words. Educated speakers use the non-standard variant in 23% of dialectal lexical items, while uneducated speakers use the non-standard variant in 83% of dialectal items. Even though the semantic factor increases the use of the non-standard variant in dialectal items, the education factor hinders the assimilation process. The same applies to common words.

Several studies have highlighted the importance of education as a social variable (e. g. Al-Wer 2002; Assiri 2008). Education as a sociolinguistic factor is a complex one. It could be an indicator of literacy in which the use of standard variant is used as the medium of instruction (Al-Wer 2002), it could be an indicator of social status since in wealthy societies such as Saudi Arabia, educated speakers are those who belong to higher social classes (Assiri 2008: 6), and it could also be a marker that shows language contact if speakers move to urban centers to continue their studies (Al-Wer 2002). Albeit different views, most studies have one conclusion: educated speakers

Table 9. Factors favouring assimilation of the definite article in Ghawareneh Arabic: level of education

education		% non-std	N
uneducated	.84	62.0	823
educated	.20	13.2	997
	.227		1820

Corrected mean .227, range 64

Table 10. Factors favouring assimilation of the definite article in Ghawareneh Arabic: intersection of education and age

education	age	% non-std	N
educated	old	18	94
	young	8	38
uneducated	old	86	342
	young	40	168
			1820

Table 11. Factors favouring assimilation of the definite article in Ghawareneh Arabic: the intersection of education and gender

education	gender	% non-std	N
educated	males	17	86
	females	10	46
uneducated	males	68	313
	females	55	197
			1820

Table 12. Factors favouring assimilation of the definite article in Ghawareneh Arabic: the intersection of education and lexical class of words

education	lexical item	% non-std	N
educated	technical	5	149
	common	13	675
	dialectical	23	173
uneducated	technical	19	47
	common	57	544
	dialectical	83	232
			1820

are more likely than uneducated speakers to use variants that are associated with standard forms or prestige.

Likewise, the findings of this study show that the education factor is significant in the variable use of the definite prefix (*ʔal*). Educated speakers use the standard variant more often. Intersecting education with gender indicates that educated male and female speakers use the standard variant more than uneducated ones. Intersecting the findings with age shows that old uneducated speakers are more likely to use the non-standard variant while young educated speakers tend to avoid using such variant.

Since the majority of Jordanian people are more likely to receive a university degree, education cannot be a reliable indicator of social status in this culture (unlike the situation described in Assiri 2008). In fact, due to the very competitive marketplace and the scarcity of jobs available, people in Jordan, in general, and the Jordan Valley, in particular, consider obtaining a university degree mandatory for both genders. For males, obtaining a university degree can guarantee a decent, well-paid job. For females, a university degree can enhance the odds of marriage (Rashad et al. 2005). Thus, it would be crucial to view the education factor in the Ghawarneh community in line with Al-Wer (2002), who views education as a possible outlet for language contact.

Due to the fact that the Jordan Valley does not have any universities, Ghawarneh move to urban centers such as Irbid to study: Yarmouk University and Jordan University for Science and Technology (in Irbid), the University of Jordan (in Amman) and others. Educated speakers, therefore, are more likely

to be in contact with dialects in which the definite prefix does not show any assimilation to non-coronal sounds. Thus, interacting with people who use the standard pattern of assimilation can be a significant factor in the difference in use between educated and uneducated speakers. For instance, in one of the interviews conducted, a speaker has indicated that once he was in Irbid city, two Ghawarneh people were using the Ghawarneh dialect on the bus. He tried to avoid talking with them so that he did not appear to belong to the peasant agricultural area. Thus, to appear as urban, modern, and educated while being in cities, some speakers tend to change their language to fit with the language spoken in urban centers. Such an example also indicates how educated Ghawarneh speakers compare their dialect with other dialects that are spoken in urban centers. The awareness of some Ghawarneh speakers of the linguistic status of the standard variant could be best described as a case of *a change from above* (Labov 1972; Meyerhoff 2010).

In addition to being subject to language contact, I argue that the use of standard language as the medium of instruction in universities and schools can lead to the use of the standard variant more than the non-standard one, especially when it comes to assimilating the definite prefix. In Jordan, it is essential for educated speakers to have mandatory courses in Standard Arabic Grammar at schools and universities; teachers highly emphasize the pronunciation of words. The same argument has been indicated earlier with regard to the coronal sound [dʒ]. However, there is no reason why such emphasis is there in the first place. Our findings show that there is a covert contrast between two perceptually similar sounds: one is a coronal while the other is not. The covert contrast between sounds affects articulation. Since speakers are not aware of the distinction between these sounds, they tend to overlap in their assimilation of the definite article; that is, in some instances, they seem not to be assimilating the definite prefix with a coronal sound. The few cases of non-assimilatory patterns highlight the contrast between these sounds. Since the coronal sound is unmarked relative to its non-coronal counterpart, speakers are using it more frequently.

The impact of education has been evident not only on the variable use of the definite prefix but also on the choice of lexical items. It has been clear that educated speakers tend to use more technical words, while uneducated speakers used more colloquial lexical items than educated ones. The choice of the lexical item and the variable that is associated with the semantic base reveals the level of change toward the standard variable. Education increases the level of awareness that overapplying the assimilation rule to non-coronal

sounds can stigmatize the speech of literate people. Moreover, the frequency of using colloquial items is hindered by the interaction of educated speakers with speakers who use relatively different forms in urban centers. Thus, I argue that if a speaker is educated, then it is discouraged for that speaker to show any dialectal feature that contradicts the rules of the standard language and the majority of the dialects spoken in Jordan.

5 Conclusion

The study has shown that the definite article in the speech of the Ghawarneh community assimilates to non-coronal sounds, as it has been argued earlier by Bani Yasin (1980). Nevertheless, assimilation is becoming less frequent. Using the standard pattern of assimilation is viewed more positively. Thus, it is likely that speakers in the region are moving toward the standard way of assimilation. In addition, the study has shown that linguistic factors can impact the assimilation of glottal consonants. The semantic content of words also limits the use of the non-standard variant in an obvious way. The education, age, and gender factors show that the non-standard variant is more likely to be used among uneducated (62%), old (46.5%), and male (40%) speakers. The sociolinguistic factors, albeit with some difference in their level of significance, meet at one juncture; there is a tendency toward moving linguistically away from the Jordan Valley to achieve some level of urbanization. Both the attitude toward life in the Jordan Valley and the degree of contact with urban centers and with other communities determine the strength of the social factor and its impact on the variable use of the definite prefix.

In addition, the study has shown that instances of a change from above and a change from below are observable in the dialect. Speakers are aware of the stigma associated with assimilation to non-coronal sounds; thus, they tend to avoid this linguistic phenomenon in urban centers. Moreover, the study shows a case of a change from below; speakers are not aware that they were extensively using assimilation with one of the non-coronal sounds, /j/, to the extent that the contrast between /j/ and /dz/ is lost. Losing the contrast between these sounds resulted in patterns in which the definite article seemed not to assimilate for coronal sounds. The unusual behavior, however, could be a matter of a covert contrast that is hard for speakers to detect perceptually. Our conclusions are at the heart of revisiting the status of /j/ across Arabic

varieties. We predict that the presence of this sound would be inevitable. Further studies are yet to confirm this prediction.

Acknowledgements

I would like to express my appreciation to many people without whom this work would not have come into being. My greatest gratitude goes to my supervisor for the current research Dr. Gerard Van Herk for his valuable and constructive comments. My sincerest thanks are to Dr. Osama Omari for his unforgettable sociolinguistic classes that have been the cornerstone of the current work. Many thanks are to Dr. Phil Branigan, Dr. Sara Mackenzie, Dr. Nicholas Welsh, Dr. Yvan Rose, Dr. Paul DeDecker, and Dr. Maureen Scheidnes for all their insightful comments on this paper. To all participants who spent their time and shared their stories during the interviews conducted, I am so much grateful for your precious time and generous hospitality. Many thanks are owed to the editor of the SKY Journal of Linguistics and the anonymous reviewers for their insightful suggestions from which I benefited a lot in improving the manuscript.

Appendix A Interview guiding topics/questions

The interviews are conducted about general topics that involve descriptions of different aspects of life that are of interest to the participants themselves.

The interviews targeted several topics. Some topics are in line with William Labov's guiding questions and topics. Others have dealt with some areas of interests for the speakers themselves in line of culture and region.

- Topic one: Life in the Jordan Valley / Farming
- Topic two: The effect of technology and social networks
- Topic three: Dreams and nightmares
- Topic four: The fasting month of Ramadan
- Topic five: Traditions in marriage and death
- Topic six: Memorable incidents at schools/universities
- Topic seven: Childhood/memorable story
- Topic eight: Making traditional meals

English illustration of some Questions

- Where did you grow up? Do you like living in the Jordan Valley? Do you think life in Irbid city is easier? What are the differences between the city and the Jordan Valley? Do you have a farm in the Jordan Valley? What kind of vegetables/fruits do you grow there? Can you describe how working on a farm is different from working in other professions? Do you think life now is better than before? How? Why do you think so? How life now is different from that of the old days? What are some environmental problems faced by your hometown?
- What is a memorable event for you? Why do you find this event memorable? Have you ever been to a wedding? Whose wedding was it? Where was it held? What sort of gifts do people buy for the bridal couple? What kind of clothes did the bride and groom wear? How is a wedding nowadays different from that of the past? Do you think that some aspects should be changed? Why do you think such aspects should be changed? What are some of the advantages of marriage?
- What do you study at university? Why do you find this major interesting? What are the most interesting topics in your major? Why? If you have a chance to change your major, what would you choose? Why would you choose that major instead?

The Arabic version

Profile questions

بداية ممكن تعرفنا عن نفسك ؟ انت مواليد اي سنة ؟ يعني تقريبا كم عمرك ؟ بتشتغل او بتدرس ؟ سكان اي منطقة بالاغوار ؟

Topic one: Life in the Jordan Valley/ Farming

كيف بتشوف الحياة بالغور هل هي سهلة او صعبة؟ هل بتشوف في اختلاف بين الحياة بين مدينة اربد والاغوار؟ شو الاختلافات الي بتشوفها بالمدينة؟ الي مش موجودة او متوفرة بالاغوار؟ شو الايجابيات الحياة بالغور وشو هية سلبياتها؟

هل بتشوف انه في تقصير بالنسبة للبنية التحتية في الاغوار؟ بسمع انه طريق الغور سبب للحوادث شو رايبك؟ مين السبب؟

في عندكو بيارة او عند احد من معارفك في منطقة الاغوار؟ شو طبيعة المواسم الي بتزرعو فيها الخضار والفواكه الي بتزرعوها في البيارة؟ ممكن تحكيلنا عن العمل بالبيارة بالتفصيل؟ هل بتشوف انه العمل في البيارة اسهل من العمل من مهن ثانية؟ هل بتشوف في اختلاف بين الحياة ايام زمان والحياة

الان ؟ شو طبيعة التغييرات الي صارت؟ هل في موقف صار حسيت انه الناس متعاونة بالاغوار اكثر من المدينة؟ طيب بالنسبة للتعاون لو عندك سيارة واحد الاشخاص طلبها منك هل بتعطي اياها؟ طيب في ناس ممكن ما تعطي شو السبب؟

Topic two: The effect of technology and social networks

الان ممكن من الاختلافات الي تغيرت عالناس مواقع التواصل الاجتماعي ؟ شو ايجابيات استخدامها وسلبياتها؟

هل بتشوف/ي انه التكنولوجيا غيرت بحياة الناس للاسهل او للاصعب؟ ممكن توضح؟

Topic three: Dreams and nightmares

هل في حلم انت حلمته غير بحياتك او بعدك بتتذكره؟ شو كان الحلم وعن ايش ؟ ممكن تحكي لي هاد الحلم غير بحياتك؟

شو رأيك بالاشخاص الي بفسرو الاحلام؟ وبيصدقوها؟ هل في حلم انت اسمعت عنه لاحد الاشخاص وحسيته غير مقنع؟

شو تفسيرك انت للحلم الي هو حلمه؟ هل بتشوف انه الاحلام بتاثر على الاشخاص وواقعهم؟ هل الاحلام بتعني اي شي بالنسبة للناس؟

Topic four: The fasting month of Ramadan

الان احنا جايين على شهر رمضان...ممكن تحكي لنا عن هاد الشهر؟ شو السبب انه الشهر مميز عند المسلمين؟

هل بتشوف انه الصيام سهل او صعب ؟ كيف بتمضي يومك خلال شهر رمضان؟ لو تحكي لنا بالتفصيل ؟ شو اكثر المسلسلات الي بتابعها بشهر رمضان؟ ممكن تحكي لي قصة واحد من المسلسلات الي بتتذكرها؟ بقلك شهر رمضان ممكن تكثر فيه المشاكل؟ شو من العادات السلبية الي الناس بتعملها خلال الشهر؟ شو بتبع شهر رمضان ؟ هل عملت عمرة ؟ شو الفرق بين العمرة والحج؟

Topic five: Traditions in marriage and death

شو رايبك بالعادات والتقاليد؟ شو في اشي بتشعره من العادات والتقاليد الي لليوم الناس بتعمله؟ مثلا بالنسبة للزواج هل بتشوف انه جزء كبير منه ماشي على العادات والتقاليد؟ في ناس بتحكي انه الزواج هاي الايام صار صعب شو الاسباب؟ هل بتشوف في مبالغاة متبعة بالعادات والتقاليد بالنسبة للزواج مثلا؟ شو العادات والتقاليد الي المجتمع بعملها في حالة الوفاة مثلا؟ ممكن تحكي لنا بالتفصيل مثلا لشخص ما يعرف عن العادات والتقاليد؟

Topic six: Memorable incidents at schools/universities

انت درست بجامعة صح ؟ كيف شفت الحياة الجامعية؟ ممكن تحكيلنا عن تخصصك بالجامعة ؟ الصعوبات الي واجهتها خلال دراستك؟ مثلا اصعب مادة واجهتها ولبش؟ شو اهمية تخصصك بالحياة العملية؟ وشو هو مستقبله الوظيفي؟ ان كان عندك الفرصة تغير تخصصك لتخصص اخر شو ممكن يكون التخصص وشو السبب؟ هل بتشوف انه الحياة الجامعة اسهل من الحياة العملية؟ ممكن تحكيلنا عن موقف صار معك بالجامعة مع دكتور او طالب بعدك الى الان بتتذكره؟ هل بتتذكر موقف سبي صار مع احد الطلاب خلا تعامله مع الدكتاترة في الجامعة؟ نفس السؤال بس موقف ايجابي مثلا صار؟

Topic seven: Childhood/memorable story

هل بتتذكر قصة من قصص الاطفال من هذيك الايام؟ شو رايك بقصة علي بابا والاربعين حرامي مثلا ؟ شو هية هاي القصة؟ هل بتشوف انه الطفولة مرحلة مميزة ؟ شو السبب؟ لو قارنت فترة الطفولة بحايتك الان مين بتشوف اسهل؟ او اجمل ؟ شو السبب؟ برايك؟ في ناس بتحكي بتمنى ارجع صغير شو السبب؟ هل انت مع هالحكي؟

References

- Aala Addin, Ibrahim. 2016. *Mardʒiʃ ʔatulaab fii ʔanaʔuu* [A student reference in Arabic grammar]. Cairo: Dar Al Kotob Al Ilmiyah.
- Abdel-Jawad, Hasan. 1981. *Lexical and phonological variation in spoken Arabic in Amman*. Philadelphia: University of Pennsylvania. (Doctoral dissertation).
- Abdel-Jawad, Hasan & Suleiman, Saleh. 1990. Lexical conditioning of phonological variation. *Language Sciences* 12(4). 291–330.
- Abu Ain, Nora. 2016. *A sociolinguistic study in Saham, Northern Jordan*. Colchester: University of Essex. (Doctoral dissertation).
- Abu-Haidar, Farida. 1991. *Christian Arabic of Baghdad*. Wiesbaden: Harrassowitz.
- Al-Khatib, Mahmud. 1988. *Sociolinguistic change in an expanding urban context: A case study of Irbid City, Jordan*. Durham: University of Durham. (Doctoral dissertation).
- Al-Qenaie, Shamlan. 2011. *Kuwaiti Arabic: A socio-phonological perspective*. Durham: University of Durham. (Doctoral dissertation).
- Al-Shatarat, Manar. 2015. *The use and social meaning of the variant [ð] among young women in Baqaa: A sociolinguistic study of Arabic in a Palestinian refugee camp in Jordan*. Salt Lake City: University of Utah. (Master's thesis).
- Al-Tamimi, Fedaa. 2001. *Phonetic and phonological variation in the speech of rural migrants in a Jordanian city*. Leeds: University of Leeds. (Doctoral dissertation).
- Al-Wer, Enam. 1991. *Phonological variation in the speech of women from three urban areas in Jordan*. Colchester: University of Essex. (Doctoral dissertation).

- 2002. Education as a speaker variable. In Rouchdy, Aleya (ed.), *Language contact and language conflict in Arabic: Variations on a sociolinguistic theme*, 41–53. London: Routledge Curzon.
- Assiri, Ahmad. 2008. *Sociolinguistic variation in Rijaal Alma, Saudi Arabia*. St. John's: Memorial University of Newfoundland.
- Bani Yasin, Raslan. 1980. *A critical and comparative study of the dialectal speech of the Ghawarna community in the Jordan Valley in Jordan*. Leeds: University of Leeds. (Doctoral dissertation).
- Chambers, J. K. 2003. *Sociolinguistic theory: Linguistic variation and its social significance*. 2nd edn. Oxford: Blackwell.
- Eckert, Penelope. 1997a. Age as a sociolinguistic variable. In Coulmas, Florian (ed.), *The handbook of sociolinguistics*, 151–167. Oxford: Blackwell.
- 1997b. The whole woman: Sex and gender differences in variation. *Language Variation and Change* 1(3). 245–267.
- Habib, Rania. 2010. Rural migration and language variation in Hims, Syria. *SKY Journal of Linguistics* 23. 61–99.
- Hayes, Bruce & Steriade, Donca. 2004. Introduction: The phonetic bases of phonological markedness. In Hayes, Bruce & Kirchner, R. & Steriade, Donca (eds.), *Phonetically based phonology*, 1–33. Cambridge: Cambridge University Press.
- Heath, Jeffery. 2002. *Jewish and Muslim dialects of Moroccan Arabic*. London: Routledge.
- Heselwood, Barry & Howard, Sara & Ranjous, R. 2011. Assimilation of /l/ to /r/ in Syrian Arabic: An electropalatographic and acoustic study. In Zaki, Hassan & Heselwood, Barry (eds.), *Instrumental studies in Arabic phonetics*, 63–98. (Current Issues in Linguistic Theory 319). Amsterdam: John Benjamins.
- Heselwood, Barry & Watson, Janet. 2013. The Arabic definite article does not assimilate. *Leeds Working Papers in Linguistics and Phonetics* 18. 34–53.
- 2015. The Arabic definite article: A synchronic and historical perspective. In Edzard, Lutz (ed.), *Arabic and Semitic linguistics contextualized: A festschrift for Jan Retsö*, vol. 1, 157–176. Wiesbaden: Harrassowitz.
- Holmes, Janet. 1992. *An introduction to sociolinguistics*. Harlow: Addison Wesley Longman.
- Horeh, Uri & Cotter, William. 2016. Current research on linguistic variation in the Arabic-speaking world. *Language and Linguistics Compass* 10(8). 370–381.
- Kenstowicz, Michael. 1994. *Phonology in generative grammar*. Cambridge: Wiley.
- Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2(2). 205–254. DOI: 10.1017/S0954394500000338.

- Lombardi, Linda. 1991. *Laryngeal features and laryngeal neutralization*. Amherst: University of Massachusetts. (Doctoral dissertation).
- McCarthy, John & Prince, Alan. 1995. Faithfulness and reduplicative identity. In *Papers in optimality theory*, 249–384. (University of Massachusetts Occasional Papers 18). Amherst: University of Massachusetts.
- Meyerhoff, Miriam. 2010. *Introducing sociolinguistics*. London: Routledge.
- Odden, David. 1986. On the role of the Obligatory Contour Principle in phonological theory. *Language* 62(2). 353–383. DOI: 10.2307/414677.
- Omari, Osama & Van Herk, Gerard. 2016. A sociophonetic study of interdental variation in spoken Jordanian Arabic. *Jordan Journal of Modern Languages and Literature* 8(2). 117–137.
- Rashad, Hoda & Osman, Magued & Roudi-Fahimi, Farzaneh. 2005. Marriage in the Arab world. (MENA Policy Briefs). Washington, DC: Population Reference Bureau. (Report).
- Rice, Karen. 2007. Markedness in phonology. In de Lacy, Paul (ed.), *The Cambridge handbook of phonology*, 79–98. Cambridge: Cambridge University Press.
- Ryding, Karin. 2005. *A reference grammar of modern Standard Arabic*. Cambridge: Cambridge University Press.
- Sankoff, David & Tagliamonte, Sali & Smith, Eric. 2005. Goldvarb X. Department of Linguistics, University of Toronto. (Technical report).
- Trudgill, Peter. 1972. Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society* 1(2). 179–195. DOI: 10.1017/S0047404500000488.
- Watson, Janet. 2002. *The phonology and morphology of Arabic*. Oxford: Oxford University Press.
- Wodak, Ruth & Benke, Gertraud. 1998. Gender as a sociolinguistic variable: New perspectives on variation studies. In Coulmas, Florian (ed.), *The handbook of sociolinguistics*, 151–167. Oxford: Blackwell.
- Woidich, Manfred & Zack, Liesbeth. 2009. The g/ġ-question in Egyptian Arabic revisited. In Al-Wer, Enam & De Jong, Rudolph (eds.), *Arabic dialectology: In honour of Clive Holes on the occasion of his sixtieth birthday*, 41–60. Leiden: Brill.
- Zhang, Yan & Tian, Feng. 2015. Study on Markedness in Linguistics. *Sino-US English Teaching* 12(9). 666–671. DOI: 10.17265/1539-8072/2015.09.004.

Contact information:

Saleem Abdelhady
Department of Linguistics
Memorial University of Newfoundland
288 Blackmarsh Rd
A1E 1T4 St. John's
Canada
e-mail: smabdelhady(at)mun(dot)ca

A new methodology for conceptual metaphor detection and formulation in corpora: A case study on a mental health corpus

Marta Coll-Florit
Universitat Oberta de Catalunya

Salvador Climent
Universitat Oberta de Catalunya

Abstract

We describe a new methodology for conceptual metaphor detection and formulation in corpora, developed within the framework of the MOMENT project for analysing mental health metaphors. We critically review state-of-the-art methods for metaphor identification in texts, highlighting their main drawbacks for metaphor analysis in large corpora, mainly practical applicability and analytical subjectivity. Our method aims at mitigating existing drawbacks on the basis of applying the following principles: (i) working hypothesis formulation and verification at the metaphorical expression detection stage; (ii) partial use of standard methods for metaphorical focus identification; (iii) use of external expert knowledge in the form of more extensive use of dictionaries and the additional use of metaphor compendia; and (iv) the implementation of strategies for conceptual metaphor formulation, including domain formulation at two levels of generalization. Satisfactory reliability test results were obtained when we tested our method for inter-annotator agreement regarding metaphor detection and formulation using texts about mental disorders as a test corpus.

Keywords: conceptual metaphor, corpus annotation, metaphor identification methods, conceptual metaphor formulation, mental health

1 Introduction

This paper presents a new methodology for annotating conceptual metaphors (CMs) in corpora. The conceptual metaphor theory, developed by Lakoff & Johnson (1980), posits that metaphor is so pervasive in ordinary daily life that our conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature (Lakoff & Johnson 1980: 3). According

to the theory, CMs are defined as a cognitive process by which a set of mappings is established between source and target conceptual-semantic domains, where the source domain (SD) is more concrete and more directly related to experience, in contrast with the target domain (TD), which is more abstract and diffuse and lacks clear delineation (Kövecses 2002: 20). As one example, it is very common to refer to ideas in relation to food (1). Note that any given CM may be linguistically instantiated in multiple metaphorical expressions (MEs), i. e., MEs are specific linguistic cases of a CM.

- (1) CM: IDEAS ARE FOOD
 ME: a. That's food for thought.
 b. I just can't swallow that claim.
 c. His idea was half-baked.

Another key notion in conceptual metaphor theory is associated with hiding and highlighting. According to Lakoff & Johnson (1980: 10), “the very systematicity that allows us to comprehend one aspect of a concept in terms of another [...] will necessarily hide other aspects of the concept”. Consequently, the use of different CMs to refer to a single domain may reveal different conceptualizations or ways of understanding the domain. This is why studying CMs is useful to detect implicit ideas and assumptions in discourse.

Discourse analysis based on conceptual metaphor theory has been implemented in numerous fields, including advertising, medicine, architecture, economics and religion (Soriano 2012: 117). Our research is conducted in the framework of a project titled *MOMENT: Metaphors of Severe Mental Disorders* (Coll-Florit et al. 2018), whose main focus is the application of conceptual metaphor theory to the mental health field. The primary goal of the MOMENT project is to contribute to a better understanding of severe mental disorders by analysing the discourse of people with those disorders and the discourse of mental health professionals. More specifically, first-person accounts by both groups, produced in Spanish and published on the Internet (in blogs, forums, etc), are analysed with the aim of detecting (1) the kind of metaphors used by these two groups when referring to mental disorders and related experiences; and (2) the kind of frames or interpretative discourses built on the use of those metaphors.

The MOMENT project involves the annotation of a large and heterogeneous corpus of mental health discourses, where the formulation of CMs – i. e., the identification and formulation of SDs and TDs for CMs

from MEs in texts – is a key component of the analysis. However, metaphor annotation in corpora is a challenging task. For example, the volume of a corpus imposes certain restrictions on manual annotation, the analysts may not always be first level experts, there is likely to be time constraints for the analysis of MEs and the corpus may be very heterogeneous, with a variety of subjects, sources and textual genres. An additional complication in our case is the adoption of a loose approach in the detection of TDs, which only need to exhibit a relationship to some aspect of mental disorders (the life of the affected people, the symptoms, the related emotions, the medication and professional intervention, etc).

To reduce the problems of time and data diversity, in a second stage of the project semiautomatic strategies of analysis will be adopted, in line with those proposed by Stefanowitsch (2006), Ogarkova & Soriano (2014) and Semino et al. (2018). However, manual annotation of a well-balanced subcorpus of equivalent texts for each group under analysis is first necessary in order to identify the prevailing set of CMs. The methodology proposed here describes this manual annotation procedure.

In this paper we focus on CMs. Nevertheless, it is important to note that two figures directly related to CMs are also annotated in our project: (1) conceptual metonymies, which represent an entity or concept in terms of another (like metaphors), but in this case relating two entities in the same conceptual domain, and (2) metaphorical similes, which are overt figurative comparisons signalled by a comparative marker.¹ This paper mainly discusses CMs, but the section describing the annotation method also briefly presents the analytical steps concerning conceptual metonymies and metaphorical similes.

Below we describe state-of-the-art methods for metaphor detection and formulation, their main drawbacks and the approach adopted in the MOMENT project to overcome those drawbacks (section 2). Next we describe the implementation of our manual annotation method, including details about the annotation steps and the documents that accompany the annotation guidelines (section 3). Finally, we provide the results of our reliability tests (section 4), briefly discuss some variables to take into account regarding the reproducibility of our method (section 5) and present our overall conclusions (section 6).

¹ Collectively we will refer to all three figures as “figurative language” and “figurative conceptualization”, while drawing the distinction between these concepts when necessary.

2 Existing methods for identifying metaphorical expressions and conceptual metaphors

The state-of-the-art methodology for ME detection in texts is the metaphor identification procedure (MIP, Pragglejaz Group 2007), subsequently refined and extended by Steen et al. (2010b) as MIPVU (where VU stands for Vrije University in Amsterdam). Steen's five-step procedure (1999; 2007) is the reference method for CMs accounting. Although it is not explicitly stated by their authors, these methods best apply to metaphorical uses of lexical words; thus they are seldom applied to detect metaphorical meanings expressed by grammatical elements.

The Pragglejaz Group (2007), in fact, focuses on identifying metaphorically used words (MUWs). The procedure is based on the distinction for all the words in the text between their meaning in that context and their so-called basic meaning. If the contextual meaning can be understood by comparison to the basic meaning, then a metaphoric use of the word in the text is determined. A reference dictionary is consulted to minimize errors and inconsistencies between analysts. Which of the dictionary meanings corresponds to the basic meaning is established by the analyst according to the following brief instruction regarding basic meanings, which "tend to be more concrete; what they evoke is easier to imagine, see, hear, feel, smell, and taste; related to bodily action; more precise (as opposed to vague); and historically older" (Pragglejaz Group 2007: 3). Note that the last condition has been ruled out by Steen et al. (2010a: 183) on the basis that older meanings are not necessarily more concrete.

By way of example, *riding* in (2)² corresponds to word sense 4 in the online version of the Macmillan Dictionary:³ "to float, or to appear to float, on water or in the air", while its basic meaning corresponds to word sense 1: "to sit on an animal, especially a horse, and control its movements as it moves along". The metaphor is meaningful in that, just as a rider controls the movements of a horse in movement, so too do the mermaids control the movements of the waves.

² The original source of this example is Steen (1999). In this and subsequent examples, the alleged MUW or metaphorical focus is boldfaced.

³ The Pragglejaz Group (2007) used the *Macmillan English Dictionary for Advanced Learners* (Rundell & Fox 2002) for their work. The online version used by us can be found at <https://www.macmillandictionary.com/> (accessed 2019-03-15).

- (2) I have seen the mermaids **riding** seawards on the waves.

MIP and MIPVU are not intended to identify the CMs underlying MUWs, although the Pragglez Group (2007: 34) does point out that the part of the procedure associated with defining basic meaning may be profitably used “to identify the source and target domains underlying metaphorical words in context”. In contrast, Steen (1999; 2007), after a first step devoted to establishing the so-called “focus” of a ME (a concept equivalent to the Pragglez Group’s MUW), does identify CMs, described in terms of propositional logic (see (7) in § 2.2.2).

In spite of their importance and widespread recognition, the Steen (1999; 2007) and Pragglez Group (2007) methodologies have several drawbacks, primarily practical applicability and subjectivity in the analyses. These will be discussed in more detail in the following sections.

2.1 The problem of time

One obvious problem with MIP and MIPVU is their application to the large-scale manual analysis of texts. The literature does not report information on the average time spent per analyst and the volume of text analysed, but it seems clear that manual analysis is not feasible for a corpus of tens of thousands of words, for which every word has to be looked up in a dictionary and then analysed for contextual and basic meanings. This problem has been tackled by some authors working with large corpora by narrowing down the range of domains to be analysed. Thus, Ogarkova & Soriano (2014) pre-selected domain-representative keywords to extract concordances using an extension of Stefanowitsch’s (2006) methodology, whereas Semino et al. (2018) combined manual with automatic semantic text labelling to filter out concordances and then selected those tagged with the most promising categories.

2.2 The problem of subjectivity

Several indeterminacy problems in metaphor analysis have been detected and analysed by Heywood et al. (2002), Semino et al. (2004), Valenzuela & Soriano (2005) and Geeraerts (2010), among others. Broadly speaking, two main difficulties are encountered: determining whether an expression is literal or metaphorical and, if metaphorical, formulating the underlying CM in terms of the appropriate SD and TD.

2.2.1 Identifying metaphorical expressions

Geeraerts (2010: 250) points out that the MIP fails to avoid the traditional difficulty of identifying semantic phenomena, namely, that “the possibility of understanding one reading in comparison with another depends on the interpreter’s ability to see the analogy – a highly subjective skill”.

Semino et al. (2004: 1277–1280) also point out that metaphoricity may be a matter of degree, as the boundary between the literal and the metaphorical is often fuzzy, as illustrated by the way the spread of cancer is talked about in terms of movement: “Although the development of cancer frequently involves the literal movement of cancerous cells inside the body (3a), the use of lexis to do with motion in our corpus often appears to be metaphorical (3b, 3c), since ‘coming back’ and ‘travelled’ do not apply literally to tumour or cancer” – we understand that because such verbs require animate agents. In any case, establishing their meaning in the text (possibly by comparison to an allegedly different literal meaning) is potentially subjective.

- (3)
- a. the way it gets there is through the blood stream.
 - b. chemotherapy can reduce the chances of things **coming back**.
 - c. it hasn’t **travelled** any more.

It is apparent that decisions concerning the literal meaning of a word can determine the metaphoricity of the phrase or sentence where that word is used. Regarding the word *freedom* in (4), for instance, Heywood et al. (2002: 46–47) suggest that its use “could be seen as metaphorical if one decided that the concept it evokes relates directly to domains such as slavery and imprisonment”, whereas “a more general interpretation of the concept of freedom as relating to the ability to choose” would point to a literal meaning.

- (4) they had so kindly offered **freedom**.

Note that, before the advent of the MIP, the focus of the debate regarding CM identification was discriminating between the figurative and literal meanings of words and expressions. However, the fact that the MIP draws a distinction between figurative and basic, rather than literal, meanings greatly simplifies the problem, as the MIP offers instructions for establishing basic meaning, whereas no instructions are available for establishing literal meanings. But, on the other hand, the MIP favours figurative readings since, while the distinction between the figurative and the literal meaning of a word is not always clear, the distinction between a figurative and a basic meaning is more straightforward.

In (4), for instance, the authors eventually decided that *freedom*, in the text, was not metaphorical since it was related to “the ability to choose what one wants to do without constraints from others”, which they considered to be the literal meaning of the word. However, an analysis in MIP terms would lean the analyst towards a metaphorical reading, since a so-called basic meaning can be found in dictionaries, e. g. “a situation where you are able to go where you want because you are not in prison” (Rundell & Fox 2002). As a consequence, following the MIP, the use of freedom in (4) would be considered metaphorical, as it contrasts with such a basic meaning.

As Geeraerts (2010: 207ff.) points out, an important source of indeterminacy is the polysemous structure of lexical units, and, at the very least, we need to consider the possible existence of dead metaphors, i. e., expressions that may be metaphorical from a diachronic point of view but have lost their metaphorical motivation for the average contemporary user. An example is (5a), which may not need to be accounted for by means of the metaphor A MOUNTAIN IS A PERSON. Otherwise, “there would just be an extension of the semasiological structure of foot, whereas the meaning of mountain could be left for what it is” (Geeraerts 2010: 208). Similar polysemy-related challenges for analysts arise when meanings originating in metaphorical extension are more frequent in contemporary language than those which could be reasonably considered literal. Semino et al. (2004: 1284–1285) report the case of *erupt* (5b) and *eruption*. Entries with those words in corpora of British English primarily reflect activities, emotions and different kinds of entities, and only reflect volcanic activity in 38% of the entries.

- (5) a. the **foot** of the mountain.
 b. something is gonna suddenly **erupt** and it’s all going to be all over.

Some cases of indeterminacy are seemingly unsolvable even considering the semasiological structure of lexical units. Heywood et al. (2002: 46) exemplify the case with (6), which seems to work both literally and metaphorically. While the expression in context can be analysed literally as being to do with the current physical location of *they*, on the other hand, “the possibility of a mapping from the domain of location to the domain of human activities in general means that the question could be to do with what has become of them”. In such cases, the decision on metaphoricity will rely on a very fine-grained interpretation of the co-text.

(6) **Where** had all they **gone**?

While aware of the complexity of these and similar problems, Ogarkova & Soriano (2014) and Semino et al. (2018) make flexible use of the MIP/MIPVU for ME identification in their analyses of large corpora, i. e., they draw on the contrast between contextual and basic meanings.

2.2.2 Formulating conceptual metaphors

The second main problem concerning subjectivity in metaphor analysis is domain labelling. Ding (2011: 72) notes that, while the MIP is a well-established procedure for identifying MUWs, it does not address CM formulation, and, furthermore, that the main problem with Steen's (1999; 2007) method is that it does not offer instructions to help in determining CM domains and correspondences.

As mentioned previously, Steen (1999; 2007) describes linguistic metaphors in terms of propositional logic, which associates concepts expressed in a text with underlying concepts. Non-expressed concepts are initially variables and then are instantiated. Steen (1999: 67) makes the metaphor underlying (7a) explicit by first formulating (7b) which, in turn, is transformed into (7c). The interesting aspect is that the instantiation of F , y and y' , i. e., identification of the concepts metaphorically underlying those present in the text, relies on the analyst's intuition or linguistic and psychological knowledge – in other words, the interpretation is subjective. The only instruction is that “prototypical or default knowledge about the source domain” (Steen 1999: 71) has to be activated.

- (7) a. I have seen the mermaids **riding** seawards on the waves.
 b. $F(\text{MERMAIDS}, \text{WAVES}) = \text{RIDE-ON}(y, y')$
 c. $\text{FLOAT}(\text{MERMAIDS}, \text{WAVES}) = \text{RIDE-ON}(\text{JOCKEY}, \text{HORSE})$

Steen's (1999) method, furthermore, does not allow clear labelling of the possible CM underlying (7a) as, by splitting the metaphoricity into three correspondences, TO FLOAT IS TO RIDE-ON, A MERMAID IS A JOCKEY and A WAVE IS A HORSE, it offers three prospects for concept labelling. Choosing labels for concepts or domains from among several possibilities is one of the most common problems in CM annotation, i. e., for a given scenario, the analyst might choose either a formulation based on the event or a formulation based on one of its arguments (Semino et al. 2004: 1276, 1281).

However, it is not only the internal structure of a certain frame, such as *to ride on horses* in (7), that can yield alternative domains for the metaphor. Domains may originate in entirely different frames, as Steen (1999: 71) and Geeraerts (2010: 207ff.) have pointed out. The latter author highlights *WAR* and *GAME* as alternative domains in cases like (8a), as *win* is commonly associated in the literature with the formulation *ARGUMENT IS WAR*. But the expression could be also perfectly consistent with *ARGUMENT IS GAME-PLAYING*, as in (8b). As Semino et al. (2004: 1284ff) have pointed out, domain selection can be biased by knowledge of extant conventional formulations.

- (8) a. I **won** the argument.
 b. lay one's cards on the table.

Ogarkova & Soriano (2014) introduced an interesting innovation by formulating CM in terms of two levels of generalization. In this way, they properly account for the theoretical distinction between generic-level and specific-level metaphors (Lakoff & Turner 1989: 80–81). In this approach, there is an inherited hierarchical structure among metaphors so that underspecified generic CMs pass on their structures to specific-level CMs. For example, *LIFE IS A JOURNEY* and *LOVE IS A JOURNEY* are specific-level metaphors of the generic *LONG-TERM PURPOSEFUL ACTIVITIES ARE JOURNEYS*. Likewise, Semino et al. (2004: 1291), while not unfolding domain formulation in levels, opt for mappings that most closely correspond to linguistic expressions.

2.3 Summary

To sum up, we list the following main problems with standard methods for identifying MEs and CMs in analyses of large corpora, of direct relevance, moreover, to our *MOMENT* project:

- The MIP/MIPVU approaches are impractical for large corpora, given the analytic detail required to determine the metaphoricity of each and every word in a corpus. One partial solution to this problem is to narrow down the volume of text under analysis, as done for corpus analysis projects by Ogarkova & Soriano (2014) and Semino et al. (2018).
- The analyst's subjectivity affects how the metaphoricity of an expression is determined. This problem is alleviated in the MIP/MIPVU

approaches by relying on the basic (as opposed to literal) meaning of a lexical unit and by making use of external expert knowledge (dictionaries). It is worth noting, however, that advancing the notion of basic instead of literal meaning favours metaphorical readings.

- No precise method for determining conceptual domains is yet available. Indetermination may stem from the need to choose a suitable frame and to choose elements (concepts) within that frame and also the level of generality of the comparison. This multiplicity of possible categories makes it difficult to achieve a reasonable level of inter-annotator agreement.

3 A methodology for detecting and formulating conceptual metaphors

To provide MOMENT project analysts with a method that mitigates the problems outlined above, we established a procedure based on the following principles:

- Working hypothesis formulation and verification at the ME detection stage.
- Partial use of standard methods for metaphorical focus identification.
- Use of external expert knowledge in the form of more extensive use of dictionaries and the additional use of metaphor compendia.
- Implementation of strategies for conceptual metaphor formulation, including domain formulation at two levels of generalization.

Below, in section 3.1, we present the rationale for these four principles and then, in 3.2, we describe their practical application via the annotation method.

3.1 Methodological principles

In this section we present in more detail the four methodological principles that form the basis of our new approach for detecting and formulating conceptual metaphors in corpora.

3.1.1 Using working hypotheses

The problem of time outlined above could be solved by reducing the volume of text to analyse manually, as done in previous corpus analysis projects. However, we did not want to rely on concordance extraction in the first phase of the project as, given the diversity of texts and subjects to be analysed (several types of people diagnosed with mental disorders and several types of mental health professionals), we did not want to be biased by keyword pre-selection.

Therefore, the following strategy was implemented: the analyst is instructed to intuitively pre-select clauses that seem to include MUWs. Their selection as hypothetical MEs can be considered a “working hypothesis”, which is later confirmed or rejected systematically by applying MIP to the words in the clause. The use of working hypotheses is a well-established practice in quantitative and qualitative research, which we apply at a micro-level. Since only the words of hypothetical MEs are analysed – and not every word as in the standard MIP approach – the time needed to apply the MIP is substantially reduced. There is obviously a risk that MEs may go undetected. However, the aim is not to detect each and every ME in the corpus but to identify predominant CMs. The possible loss of some MEs is offset by significant time savings, to the point of ensuring the actual feasibility of the project.

Lastly, one of the specifications to analysts is that a ME should be selected as hypothetical only if it applies to severe mental disorders in accordance with a list of corresponding thematic fields. This specification makes it possible to dismiss a large number of CMs that Geeraerts (2010) would deem problematic as being very conventional or possibly dead, as illustrated in (5a).

3.1.2 Partial application of the MIP

Despite the problems implied by decisions about metaphoricality (see § 2.2.1), like Ogarkova & Soriano (2014) and Semino et al. (2018) we opted to use the MIP, as its advantages outweigh its disadvantages. Even though it favours metaphorical readings and so incurs the risk of overanalysis (because it relies on basic rather than literal meanings), it is well established and provides clear instructions for carrying out analyses at the lexical level.

However, our use of the MIP is partial in that it is not applied to the whole text but only to clauses selected as hypothetical MEs. It is also partial in that we exclude the historicity criterion, as per Steen et al. (2010b).

3.1.3 Using compendia

In MOMENT, the analysts use existing compendia of metaphors or compendia specifically compiled for the project. The compendia are used in the first instance to formulate the CM, and only when the ME does not fit any of the available models are the domains inferred by the analysts.

While the risk exists (as mentioned above) that prior knowledge may bias the analysis, this is always the case when resorting to external information – as done in the MIP when dictionaries are used to determine word meanings. Both dictionaries and compendia are used because they are regarded as expert knowledge resources (the information they contain is viewed more as genuinely useful than as risky). All things considered, the risk of bias is outweighed by the benefits: inter-annotator agreement is facilitated thanks to the mitigation of some of the indeterminacy and subjectivity factors (as discussed in § 2.2.2), results consistent with previous research findings are obtained and the annotator’s task is facilitated and abbreviated by the availability of expert knowledge.

3.1.4 Strategies for conceptual metaphor formulation

We have developed strategies to formulate CMs (i. e., inferring SDs and TDs) when the ME does not seem to fit any of the CMs in the compendia. These strategies are explained to the analysts in the annotation guidelines. We regarded this as necessary for two reasons: firstly, as reported above (§ 2.2.2), as far as we know no instructions have yet been developed for this task, outside of Steen’s suggestion to base inferencing on “prototypical or default knowledge about the source domain” (1999: 71); secondly, better inter-annotator agreement is more likely when a common approach is used for decision-making. Thus, one of the main innovations of our method is the description of systematic strategies for conceptual metaphor formulation to maximize agreement between analysts.

Our strategies are based on substituting words in the ME with other words affording a literal reading. When a verb is used figuratively, we substitute one of its arguments with a more prototypical word (as suggested by Steen 1999) in order to reach a literal reading. In other cases the MUW is substituted by a key concept extracted from the dictionary definition of the word. This is in line with the Praggeljaz Group (2007: 34), which suggests that “Metaphor scholars [...] may profitably use the MIP, especially the step associated with

defining basic meaning, to identify the source and target domains underlying metaphorical words in context”.

Whenever possible, CMs are formulated at two levels of generalization, in line with Ogarkova & Soriano (2014). This approach will be helpful at a later interpretative stage to distinguish between a specific level closer to the field under analysis (in our case, severe mental disorders) and an abstract level more useful for drawing generalizations about broad types of metaphors used in the texts. Moreover, as pointed out by Grady et al. (1999, as cited in Semino et al. 2004: 1291) “conceptual domains are often too general as units of analysis for conceptual metaphors and [...] many mappings are better described as associations between source and target concepts, belonging to distinct domains”.

Our four strategies for CM domain inferencing are described in what follows.

A. Substitution by a prototype argument of a metaphorically used verb

When the metaphorical focus is a verb and its selectional preferences appear to be violated, the analyst determines the prototypical argument(s) for the verb and checks whether substitution of (any of) the argument(s) in the text by the prototype will make the semantic incongruence disappear. If yes, the prototypical argument is established as the SD and the word from the text is established as the TD. As an illustrative example, in applying the MIP to (9a), with *emancipado* (‘emancipated’) detected as the metaphorical focus of the ME, the reasoning is as follows. Using *dolor* (‘pain’) as a prepositional phrase/genitive argument of *emancipar* would violate the verb’s selectional preferences. Since a suitable prototypical argument, inferred from the dictionary definition of *emancipar* (9b), is *autoridad* (‘authority’), using *autoridad* rather than *dolor* in the text would meet the verb’s selectional preferences. Consequently, the CM is formulated using *dolor* as the TD and *autoridad* as the SD (9c).⁴

⁴ In the conceptual metaphor theory literature, concepts and domains are conventionally represented in uppercase. The same convention is used by us, but in example’s explanations, for stylistic reasons, we use lowercase for the word which evokes the concept.

- (9) a. *no hay nadie **emancipado** del dolor.*
 ‘nobody is **emancipated** from pain.’
- b. Emancipar 1: Liberar de la autoridad legal paterna, de la tutela, de la servidumbre o de otro tipo de subordinación o dependencia
 ‘Emancipate 1: To grant freedom from parental legal authority, guardianship, servitude or another type of subordination or dependence’.
- c. EL DOLOR ES UNA AUTORIDAD
 PAIN IS AN AUTHORITY

The wording of the argument will preferably be extracted from the dictionary definition of the focus (for example *autoridad* from the dictionary definition of *emancipar*). Note that, in this strategy, while domains may be generated from the focus, the focus itself is not established as a domain of the underlying metaphor (e. g. *emancipar* itself is not considered a source or a target domain).

B. Substitution by keywords from the dictionary definition of the focus in a lexicalized metaphor

When the metaphorical focus is not a verb, or when it is a verb but strategy A is not applicable, the analyst formulates an operational comparison between the contextual and basic meanings of the focus. If the contextual meaning of the focus is lexicalized (i. e., it corresponds to one of the meanings of the focus in the reference dictionary), the metaphor formulation is drawn from the contextual and basic dictionary definitions. A word representing the contextual meaning is annotated as the TD, and a word corresponding to the basic meaning is annotated as the SD.

In (10a), *estigma* is detected as the focus on applying the MIP. The contextual meaning corresponds to word sense 2 (10b), whereas the basic meaning (10c) is more concrete, more visible and related to the body. Therefore, the CM is formulated using *deshonra* (‘dishonour’) as the TD and *marca en el cuerpo* (‘mark on the body’) as the SD (10d), both chosen as the more explanatory lexical units in the dictionary definitions. As in strategy A before, the focus itself (*estigma*) is not established as a domain of the underlying metaphor.

- (10) a. *Hace más de un año que apareció el **estigma**.*
 ‘The **stigma** appeared more than a year ago.’
- b. Estigma 2: Motivo de deshonra o de mala fama.
 ‘Stigma 2: Reason for dishonour or bad reputation.’

- c. Estigma 1: Marca o señal en el cuerpo.
'Stigma 1: Mark or scar on the body.'
- d. LA DESHONRA ES UNA MARCA EN EL CUERPO
DISHONOUR IS A MARK ON THE BODY

C. Substitution by keywords from the dictionary definition of the focus in a non-lexicalized metaphor

If the contextual meaning is not lexicalized (i. e., it does not correspond to any of the meanings of the focus in the reference dictionary), the analyst will infer meaning from the context or co-text. The word representing the contextual meaning of the focus is then annotated as the TD, and the focus itself or a label corresponding to its basic meaning is annotated as the SD.

On applying the MIP in (11a), *pegatina*, an informal synonym for *adhesivo* ('sticker'), is detected as the focus. From prior context it is inferred that this word refers metaphorically (and pejoratively) to a person's diagnosis (*diagnóstico* in Spanish), whereas its only meaning in the dictionary is (11b). The CM is thus formulated using *diagnóstico* as the TD and *adhesivo* as the SD (11c), since the more formal label extracted from the dictionary definition is preferred.

- (11) a. *Tomé la decisión de quitarme la **pegatina**.*
'I made the decision to get rid of my **sticker**.'
- b. Pegatina 1: Adhesivo pequeño que lleva impreso un texto o una imagen.
'Sticker 1: Small adhesive piece of paper with text or pictures printed on it.'
- c. UN DIAGNÓSTICO ES UN ADHESIVO
A DIAGNOSIS IS A STICKER

D. Substitution and adscription to a more general mapping

The strategies presented above generate metaphorical comparisons at the specific level. But in some cases the comparisons may be more general, leading to the formulation of appropriate general-level metaphors. In other cases the comparison is recognized as an ontological correspondence for a CM (Lakoff & Johnson 1980), inasmuch as one can think of other mappings between concepts in the domains. These two cases of generalization correspond to the two cases of specification of general metaphors posited by Ogarkova & Soriano (2014), which they call "special case" and "entailment",

respectively. Here we present an example of the latter.

In (12a), *apostamos* is one of the expression's MUWs. Following the above-described strategy B, the CM is initially formulated as (12b). However, the basic meaning of *apostar* (12c) in the reference dictionary indicates that the concept belongs to the gambling domain. Several elements can be distinguished within this complex frame: gambling, gaming houses, money, getting rich, going bankrupt, etc. Therefore, the mapping in (12b) can be reasonably regarded as one of several possible mappings in a more general analogy between gambling and life.⁵ Lakoff & Johnson (1980: 51), in fact, postulate LIFE IS A GAMBLING GAME as a conventional metaphor that covers expressions such as *I'll take my chances, the odds are against me, I've got an ace up my sleeve*, etc. Since equivalents for these and similar expressions exist in Spanish, it seems reasonable to formulate a more general metaphor as in (12d). The fact that the CM underlying *apostamos* is richly annotated at two levels – specific in (12b) and generic in (12d) – captures both aspects of the conceptual mapping.

- (12) a. *Elegimos un camino peligroso, **apostamos** fuerte, alto...*
 ‘We choose a dangerous road, we **bet** hard, high...’
- b. ARRIESGAR ES APOSTAR
 TO RISK IS TO BET
- c. Apostar 2: Referido a una cantidad de dinero, arriesgarla para poder participar en el juego que consiste en acertar el resultado de algo, de forma que, si se acierta, se recibe una cantidad de dinero mucho mayor. ‘Bet 2: Referring to a quantity of money, risking it to participate in the game that consists of getting the result of something right, so that, if right, a much larger amount of money is received.’
- d. LA VIDA ES UN JUEGO DE AZAR
 LIFE IS A GAMBLING GAME

3.2 Annotation method

The method for manually annotating figurative expressions in the MOMENT project is organized along mutually exclusive paths, consisting of steps with explicit instructions for detecting and formulating three possible conceptual figures: CMs, conceptual metonymies and metaphorical similes. Although in this paper we focus on CMs, below we also briefly describe the analytical steps

⁵ The concept LIFE IS chosen as TD because the co-text indicates that the speaker is talking about his life trajectory.

concerning conceptual metonymies and metaphorical similes. The overall structure of the process is depicted in Figure 1.

Annotators are issued a document of annotation guidelines and several complementary documents. The former details the procedure described below and the latter are the following:

1. List and description of thematic fields specific to severe mental disorders
2. List of lexical units excluded as potential MUWs
3. List of comparison markers for simile detection
4. Compendium of mental health CMs
5. Compendium of mental health conceptual metonymies
6. Compendium of general purpose CMs
7. Compendium of general purpose conceptual metonymies.

Documents 1–5 were specifically compiled for the project. Document 1 has been drawn up based on the authors' previous work on schizophrenia metaphors (Climent & Coll-Florit 2017; Coll-Florit et al. 2019); it lists and briefly describes semantic fields considered relevant for the analysis of discourse in mental disorders (the life of the affected people, the symptoms, the related emotions, the medication and professional intervention, social prejudices and discrimination, etc). Document 2 lists certain word classes, delexicalized words and data-specific terms to be excluded from the analysis, as per Semino et al. (2018: 59).⁶ Document 3 is based on the Real Academia Española (2009: 3408–3420) grammar of the Spanish language. Documents 4 and 5, compiled from Barcelona (1986), Semino (2008: 178–190), Climent & Coll-Florit (2017) and Coll-Florit et al. (2019), consist of metaphor formulations with representative examples. Document 6 is the Master Metaphor List (Lakoff et al. 1991). Document 7 was extracted

⁶ These are the following: (1) terms very commonly used in the field such as *brote* ('flare-up', literally 'sprout'), which were metaphorical in origin but have become medical terminology; (2) prepositions when carrying no semantic content, with some exceptions, such as the locative uses of *en* ('in'); (3) lexical verbs functioning as auxiliary or modal markers such as *tener* ('have') or *acabar* (literally 'bring to an end'); and (4) interjections, e. g. *qué diablos* ('what the hell'), as they can be analysed metaphorically but have become extremely conventional.

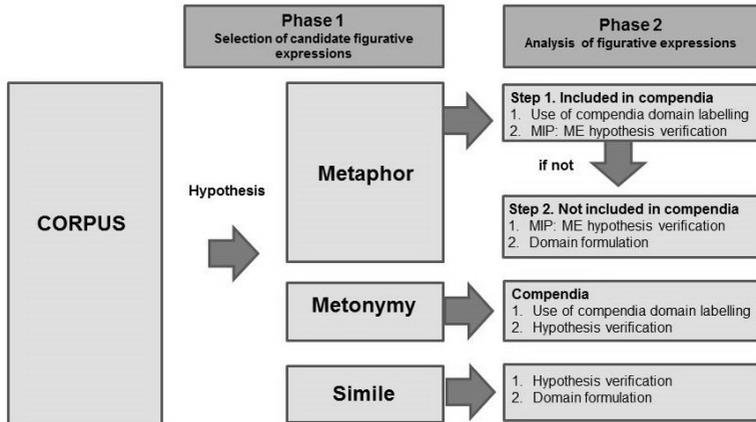


Figure 1. Overall structure of the annotation method

from Littlemore & Tagg (2016) (who drew, in turn, on Radden & Kövecses 1999).

The analysts use two Spanish language dictionaries, CLAVE (Maldonado 2012) and *Diccionario de la Lengua Española* [DLE] (Real Academia Española 2001), chosen for the following reasons: both are online so lookup is facilitated; CLAVE is a recent usage dictionary with clear-cut word meanings and illustrative examples; and DLE is the standard Spanish language reference dictionary. Instructions are to preferably use CLAVE and to resort to DLE in cases of doubt.

The method consists of two main phases: selecting the hypothetical figurative expression and analysing the hypothetical figurative expression. Sentences are randomly presented to the analyst. While the immediate context is undeniably primordial, each sentence is presented with the immediately preceding and subsequent sentences, and analysts are also instructed to consult the original text (easily accessed through hypertext links) if doubts remain.

3.2.1 Phase 1: Selection of candidate figurative expressions

The analyst reads the sentence and its context to capture the general meaning and, checking against the list of thematic fields for severe mental disorders, decides whether the sentence contains one or more candidate figurative expressions (those perceived not to belong are rejected). The following general clues are used to hypothesize the occurrence of figurative expressions and to distinguish between metaphors, metonymies and similes:

- **CMs.** An indirect or non-literal use is made of a word or a group of words that, in the context, seems to express some kind of comparison or resemblance between concepts in such a way as to make the discourse more expressive or understandable.
- **Conceptual metonymies.** A noun or a noun phrase represents and/or replaces another noun or noun phrase belonging to the same domain of knowledge and the concepts represented are related by a spatial, temporal, causal or part-whole contiguity.
- **Metaphorical similes.** The sentence contains a comparison marker.

Hypothetical figurative expressions are extracted as clauses from the sentence with enough immediate context to be understandable.

3.2.2 Phase 2: Analysis of figurative expressions

The analysis evolves along separate no-return paths referring to the three kinds of conceptual figures postulated by the analysts on the basis of hypotheses.

Conceptual metaphor analysis

CMs are annotated in one of two ways, depending on whether the metaphor is included or not included in compendia (Step 1 or Steps 1 and 2, respectively).

Step 1: Metaphors included in compendia

In this step, in the first place the CM domains are provisionally annotated and, in the second place, expressions are analysed in order to verify or reject both their status as an ME and their correspondence to the tentative CM. The process is as follows: for each hypothetical ME, the analyst checks the

metaphor compendia to determine if a corresponding suitable CM exists. Analysts are instructed to first check the compendium of mental health metaphors and only use the general purpose compendium if the metaphor is not found in the former. If a suitable CM is found in the compendia, this tentative CM is used (i. e. domains are annotated) at the generic level; otherwise the analyst proceeds to Step 2. The specific-level formulation is determined from the text – in a straightforward way if both domains are explicit or applying one of the domain inference strategies (as described in § 3.1.4 above). Last but not least, the metaphoricity of the hypothetical ME is verified or rejected by application of the MIP to all its lexical units. If at least one of these units is judged to be an MUW, the metaphoricity of the ME is verified; and if the ME is judged to be congruent with the tentatively formulated CM, then the annotation is deemed valid. Otherwise, the provisional annotations are cancelled, and the process goes to Step 2.

As can be seen, this step proceeds in reverse order to what could be expected, since determining if an expression is metaphorical typically comes before trying to label the corresponding CM. Our experience in metaphor annotation has shown that dealing first with intuitions based on previous work (compendia) and then verifying or counter-verifying them results in more consistent labellings and annotation time saving. This occurs mainly because the same generic CMs included in compendia are being repeated in texts: e. g. *THE DISORDER IS A LOCATION* is instantiated in several specific locations: hell, pit, hole, gutter, labyrinth, etc. Therefore, what differentiates this approach from past practices of CM accounting (e. g. Steen 1999; 2007) is that our method is based on metaphor compendia in the first instance.

As an example of Step 1, in (13a) the analyst selects the clause *conocer el infierno* (13a) as an hypothetical ME, with *infierno* as the potential focus. The mental health metaphor compendium includes *EL TRASTORNO ES UN LUGAR* ('*THE DISORDER IS A LOCATION*'), considered applicable to the expression. Thus, at the generic level, *trastorno* is annotated as the TD and *lugar* as the SD. As for the specific level, in coherence with the text, *infierno* ('HELL') is chosen as the SD and the TD results from hypothesis verification. The analyst judges that the contrast between meanings 8 (13b) and 1 (13c) for *infierno* in the DLE dictionary is metaphorical in accordance with MIP specifications, and that this contrast is congruent with the provisionally postulated metaphor, and so the hypothesis is verified. Moreover, applying the second domain inference strategy, i. e., substitution based on keywords from the dictionary definition of the focus in a lexicalized metaphor (§ 3.1.4), the specific TD is

inferred by comparing the two dictionary definitions. From (13b) *sufrimiento* is considered to be the more faithful representation of the meaning conveyed by the text and so the CM underlying (13a) is formulated as (13d).

- (13)
- a. *Lo que no sabía [...] es que a los pocos días iba a conocer el **infierno***
‘I didn’t know that in a few days I was going to be in **hell**’
 - b. Infierno 8: Lugar o situación que causa gran sufrimiento o malestar.
‘Hell 8: A place or situation causing extreme suffering or distress’.
 - c. Infierno 1: En la doctrina tradicional cristiana, lugar donde los condenados sufren, después de la muerte, castigo eterno.
‘Hell 1: In traditional Christian doctrine, the place where the damned suffer eternal punishment after death.’
 - d. Specific level: EL SUFRIMIENTO ES EL INFIERNO (SUFFERING IS HELL).
Generic level: EL TRASTORNO ES UN LUGAR (THE DISORDER IS A LOCATION).

Step 2: Metaphors not included in compendia

When a compendium metaphor is not identified, the procedure is reversed, i. e., the focus of the ME is first detected and the CM domains are then inferred. The analyst first applies the MIP to all the words in the hypothetical ME. If a word is eligible, it is annotated as the focus and the hypothesis is verified; the analyst then formulates the domains of the underlying CM. If no MUWs are detected by the MIP, the hypothesis is rejected and the process ends. Conceptual domain labelling consists of an initial formulation at the specific level – by application of the appropriate domain inference strategy (see § 3.1.4) – followed by a generic level annotation according to one of the following possibilities: the specific metaphor is either a subtype of a more abstract metaphor (as in 13) or is an ontological correspondence of a CM (as in 12). If a generic-level labelling is not found to be possible, the ME remains just annotated at the specific level. This is the case with examples (10) and (11), which can be regarded as *one-shot metaphors* (Steen 1999: 58–59).

Conceptual metonymy analysis

Analogously to Step 1 for CMs, the analysis consists of first determining and then verifying the metonymy. In line with Brdar (2018), who states that there are no new metonymies only new instances of known types created by analogy, it is assumed that the general compendium of conceptual metonymies

provides comprehensive coverage of all possible types at the generic level. The analyst determines whether the source term (in the text) and the target concept instantiate a compendium metonymy. If so, the source term is annotated as the focus, the compendium metonymy is provisionally annotated at the generic level and the concepts or domains directly emerging from the text are annotated at the specific level. Otherwise, the hypothesis is rejected and the process ends. The hypothesis is verified by analysing the metonymic focus using a method adapted from (Biernacka 2013: 117, cf. Littlemore & Tagg 2016). The procedure is similar to the MIP, except that, in this case, concepts are related by spatial, temporal, causal or part-whole contiguity.

By way of an example, in (14), *mundo* is postulated as the focus and is determined, in this context, to instantiate (14b), a metonymy included in the general compendium. Consequently, at the generic level *lugar* is annotated as the SD and *habitantes de un lugar* as the TD. To determine the specific level, the dictionary meanings of *mundo* are checked. Both relevant meanings are lexicalized, (14c) as the basic concept and SD and (14d) as the TD. From the definitions, the metonymy is formulated at the specific level as in (14e).

- (14) a. *Es el mundo, que no está preparado.*
 ‘It is the **world**, which is not ready yet’.
- b. EL LUGAR POR LOS HABITANTES DE UN LUGAR
 ‘THE PLACE FOR INHABITANTS’
- c. Mundo 3: Planeta o astro, esp. referido a la Tierra.
 ‘World 3: Planet or heavenly body, esp. the Earth.’
- d. Mundo 4: Conjunto o sociedad de los seres humanos.
 ‘World 4: All people or society’.
- e. EL PLANETA TIERRA POR LA SOCIEDAD
 PLANET EARTH FOR SOCIETY

Metaphorical simile analysis

Following Semino et al. (2018: 282), we understand similes as “the explicit linguistic formulation of a comparison between two unlike entities, usually signalled by expressions such as ‘like’ or ‘as if’, as in ‘Cancer is like a journey’”. In similes there is no contrast between basic and contextual meanings of the source, therefore it differs from CMs in that in this case language is used “directly” (Steen et al. 2010b: 11). Our annotation guidelines consider a simile to be metaphorical if it establishes a comparison between a target abstract term and a source concrete term, while it is considered

non-metaphorical if it just establishes some kind of similitude between entities.

Taking a comparison connector as indicating a hypothetical conceptual simile, the analyst first determines whether the contrast between the two terms is metaphorical. If so, the hypothesis is verified and the first term of the comparison is provisionally annotated as the TD and the second term as the SD. If the comparison between the two terms is literal, then the process ends. Finally, if applicable, the generic level is established as in CMs.

For instance, in (15) the presence of a simile is indicated by *como* ('like'), which reflects a comparison of the concepts denoted by *enfermo mental* and *reloj*. The analyst judges the contrast to be metaphorical, given that it is asymmetrical, highlights a single property and contains an explanation or elaboration. The concept of *la persona diagnosticada*⁷ is annotated as the TD and *reloj* is annotated as the SD (15b). Finally, the conventional metaphor (15c) is formulated at the generic level, on the basis that the comparison in the text is judged to be an instance of it.

- (15) a. *El enfermo mental [...] como un reloj debe marcar las horas.*
 'The mental patient [...] like a clock has to mark time'.
 b. LA PERSONA DIAGNOSTICADA ES UN RELOJ
 'THE DIAGNOSED PERSON IS A CLOCK'
 c. UNA PERSONA ES UNA MÁQUINA
 'A PERSON IS A MACHINE'

4 Reliability testing

To assess the validity of the method and to empirically assess its reliability, inter-annotator agreement was tested. Details of the test corpus, the analysts and the coding procedure are described below, followed by the results, reflecting the degree of statistical confidence.

4.1 Test corpus

The research was conducted in the framework of a project whose primary focus, as mentioned earlier, is the analysis of mental health metaphors. The corpus used for inter-annotator agreement testing included Spanish texts (4,143 words in 200 sentences) produced by two subjects with a severe mental

⁷ In the MOMENT project we use this term to refer to the people with a mental health condition instead of *enfermo mental* or similar derogatory or disrespectful words.

disorder, published in a book describing personal stories and experiences of the mental disorders (*El libro de Radio Nikosia*; Vv. Aa 2005).

4.2 Analysts

Corpus annotation was independently performed by the authors of the annotation method described in this paper, both linguists and both experts in conceptual metaphor theory, who have worked together in the area of mental health metaphors for four years. Before testing, the analysts, thoroughly versed in the annotation procedure, had discussed the method in several meetings.

4.3 Annotation interface and analysed items

Metaphors were coded in ten interface fields in Excel files (Table 1).

The reliability test examined inter-annotator agreement for the following main items:

- a) Identification of metaphorical language use. For the 200 sentences in the test corpus, we calculated agreement between the analysts regarding whether a sentence included metaphorical language reflecting the severe mental disorders field.
- b) Use of metaphor compendia. To assess the usefulness of the metaphor compendia as expert knowledge, we calculated agreement between the analysts regarding sentences including metaphors included in compendia.
- c) MUW annotation. For sentences that both analysts hypothesized as metaphorical, we calculated their agreement in coding the MUWs. It is important to note that this is the only item evaluated in reliability testing for MIP (Pragglejaz Group 2007) and MIPVU (Steen et al. 2010b).
- d) Conceptual domain formulation. For the identified metaphors, we calculated agreement between the analysts in the formulation of conceptual domains (SD and TD). This item has not been previously evaluated in reliability tests of metaphor annotation. Although Ogarkova & Soriano (2014) report agreement results on metaphor labelling, they are actually based on a closed inventory of CMs into which to try and classify MEs. Therefore, technically speaking, it can be regarded as metaphor classification, inasmuch as there is no application of strategies for domain formulation.

Table 1. Annotation interface for the inter-annotator agreement test

Field	Value	Description
ID-String	Natural number (1)	Identifier for each sentence in the corpus.
ID-Line	Rational number (1.1)	Identifier for each line in the Excel file. A sentence is split in several lines if it contains more than one ME.
ID-Author	Natural number (1)	Identifier for the author of the text.
String	Text	Sentence. It can be repeated as many times as a ME is detected.
Metaphorical expression (ME)	Text / NO	Clause for which metaphor use is hypothesized (in the severe mental disorder field). Standing alone it must be meaningful, although it may be discontinuous in relation to the original text. If there is no metaphorical language, this field is coded as NO.
Figure	MF, MN, SI	Figure descriptor. MF: metaphor MN: metonymy SI: simile
Focus	Text	MUWs in the ME.
Conceptual domains	Text	Formalization of the conceptual domains for the metaphor: specific and generic.
Compendia	C, N	Origin of the figure. C: included in metaphor compendia. N: not included in metaphor compendia.
Comments	Text	Open field.

Table 2. Inter-annotator agreement and statistical confidence scores

Items	Agreement	Cohen's Kappa
Identification of metaphorical language use referring to severe mental disorders	91%	0.81
Identification of metaphors included in compendia	87%	0.70
MUW annotation	97.6%	0.79

4.4 Reliability results

In this section we present the results of the agreement test between analysts before discussion. To investigate possible bias in the individual analysts, we computed the kappa test statistic, the reliability measure most widely used in previous studies of metaphorical language annotation (Markert & Nissim 2003; Pragglejazz Group 2007; Steen et al. 2010b). More specifically, we used Cohen's kappa, given that we were testing agreement between a pair of analysts. In interpreting kappa values, the literature indicates that values in the range 0.61–0.80 point to substantial agreement.

Our test results (Table 2) point to a high degree of agreement between the analysts, at 91% for the identification of metaphorical language use related to the semantic fields of severe mental disorders (kappa value 0.81), and at 87% for the identification of metaphors included in compendia (kappa value 0.70). Regarding the annotation of MUWs, the percentage of agreement rises to 97.6% (kappa value 0.79). Taking kappa values reported for previous MIP and MIPVU inter-annotator agreement tests as the gold standard in MUW annotation, the Pragglejazz Group (2007) reported kappa values of 0.62 for conversation texts and of 0.72 for news texts (MIP), while Steen et al. (2010b) reported kappa values ranging from 0.70 to 0.96 for different text genres (academic texts, fiction, news and conversation), with the conversation genres obtaining the lowest values (MIPVU). Our kappa value results, therefore, are better than those for MIP and similar to those for MIPVU.

Finally, Table 3 shows the level of agreement in the formulation of conceptual domains, a concept not previously evaluated. For this analysis, we considered fragments for which both analysts identified metaphorical language use. Since conceptual domains are not a closed tagset, the categories are not mutually exclusive as required for kappa calculations and,

Table 3. Inter-annotator agreement in conceptual domain formulation

Both SD and TD	Agreement		
	At least one domain	SD	TD
71%	87%	78%	80%

consequently, we only report the percentage agreement between analysts. Note that two formulations for the same domain were considered equivalent if there was agreement at either the specific or the generic level. Additionally, two formulations were recognized as equivalent if they used synonyms or different grammatical categories but clearly referred to the same entity or concept (e. g. in Spanish *transferencia* (n) / *transferir* (v) [‘transfer’]).

As can be observed in Table 3, the analysts proposed the same formulation for both metaphor domains in 71% of the cases and agreed regarding at least one domain in 87% of the cases. Interestingly, there was no substantial difference regarding agreement as to the SD (78%) and the TD (80%).

Finally, it is worth mentioning that, when the annotation includes domain formulation at two levels of generalization, there is agreement in both levels in 76% of the cases. Furthermore, when there is agreement in only one level (in the 24% remaining cases), we did not find significant differences in agreement between the specific and the generic levels.

5 Reproducibility of the method

The new methodological approach presented in this paper offers systematicity for both ME detection and CM formulation, which favours its replicability to other corpora. However, there are still some variables that could introduce noise in the reproducibility of the method.

Regarding ME detection, different analysts may choose different basic and contextual dictionary definitions. In a related way, some definitions are ambiguous between a basic and a figurative meaning. Nevertheless, note that these are not new limitations but inherent to the application of the MIP. With reference to CM formulation, different researchers may choose different keywords from dictionary definitions when applying strategies for domain formulation. Therefore, some degree of subjectivity can still be expected in applying the method, although this is a limitation which is unavoidable when dealing with manual annotation of corpora.

6 Conclusions

In the context of research conducted in the framework of MOMENT, a project that aims to analyse a large corpus of mental health texts in the light of conceptual metaphor theory, we have developed a feasible and reliable methodology for detecting MEs in discourse and establishing the underlying cross-domain mappings.

In designing our method, we fundamentally pursued two aims: (i) to maximize agreement between analysts by reducing the impact of subjectivity and indetermination as much as possible, given that conceptual domain inference is a highly elusive task for which standard methods offer little guidance; and (ii) to render the task feasible in terms of time. Our method is structured as a workflow, with mutually exclusive paths, designed on the basis of the following principles: use of working hypotheses, partial use of the MIP (Pragglejaz Group 2007) to establish metaphoricity, use of external expert knowledge (dictionaries and compendia) and the implementation of specific strategies to infer underlying conceptual domains.

Thus, this method presents some innovations and makes new contributions both in metaphor detection and formulation. In regard to detection, the use of working hypotheses at the ME stage substantially reduces the time needed to apply the MIP, so that it does not have to be applied to each and every word in a corpus. In relation to formulation, the use of metaphor compendia in the first instance facilitates and abbreviates the annotator's task and fosters inter-annotator agreement. Moreover, and crucially, this method presents several systematic strategies for conceptual metaphor formulation, based on a more intensive use of dictionaries. This last point is especially relevant taking into account that the field of metaphor studies lacks a standardized method for conceptual metaphor formulation.

Satisfactory results were obtained when we tested our method for inter-annotator agreement with a 4,143-word test corpus. Regarding ME detection, the statistical confidence values were better than those reported for the MIP (Pragglejaz Group 2007) and similar to those reported for the derivative MIPVU (Steen et al. 2010b). With respect to domain formulation, agreement was reached in around three quarters of the cases, a satisfactory result considering that conceptual domains are not a closed tagset but are open to the annotator's subjectivity. Therefore, we have assessed the validity of this new methodology that, although applied to a corpus of mental health texts in this paper, can be easily applied to essentially any corpus.

Acknowledgements

This research was conducted in the framework of the project *MOMENT: Metaphors of severe mental disorders. Discourse analysis of affected people and mental health professionals*, funded by the Spanish National Research Agency (Agencia Estatal de Investigación, AEI) and the European Regional Development Fund (ERDF), within the National Programme for Research Aimed at the Challenges of Society. Ref. FFI2017-86969-R (AEI/ERDF, EU).

We would like to thank the two anonymous reviewers whose comments and suggestions greatly helped to improve this manuscript. All remaining flaws are ours.

Abbreviations

CM	conceptual metaphor
ME	metaphorical expression
MIP	metaphor identification procedure
MIPVU	metaphor identification procedure Vrije University
MUW	metaphorically used words
SD	source domain
TD	target domain

References

- Barcelona, Antonio. 1986. On the concept of depression in American English: A cognitive approach. *Revista Canaria de Estudios Ingleses* 12. 7–33.
- Biernacka, Ewa. 2013. *The role of metonymy in political discourse*. Milton Keynes: The Open University. (Doctoral dissertation).
- Brdar, Mario. 2018. On the novelty of novel metaphors and metonymies. (Plenary lecture at AELCO/SCOLA 2018, 11th International Conference of the Spanish Cognitive Linguistics Association, 19 October 2018).
- Climent, Salvador & Coll-Florit, Marta. 2017. La metáfora conceptual en el discurso psiquiátrico sobre la esquizofrenia [Conceptual metaphor in psychiatric discourse on schizophrenia]. *Ibérica* 34. 187–208. (Special Issue on Metaphor in Science and Technology).

- Coll-Florit, Marta & Climent, Salvador & Correa-Urquiza, Martín & Hernández, Eulàlia & Oliver, Antoni & Pié, Asun. 2018. MOMENT: Metáforas del trastorno mental grave: Análisis del discurso de personas afectadas y profesionales de la salud mental [MOMENT: Metaphors of severe mental disorder: Discourse analysis of affected people and mental health professionals]. *Procesamiento del Lenguaje Natural* 61. 139–142.
- Coll-Florit, Marta & Miranda, Xavier & Climent, Salvador. 2019. Metáforas de la esquizofrenia: Un estudio sobre el discurso de afectados y profesionales [Metaphors of schizophrenia: Discourse analysis of sufferers and professionals]. *Revista Española de Lingüística Aplicada (RESLA) / Spanish Journal of Applied Linguistics (SJAL)* 32(1). 1–31.
- Ding, Yan. 2011. *Metaphors of sadness: Intraconceptual and interconceptual variation*. Hong Kong: University of Hong Kong. (Doctoral dissertation).
- Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford: Oxford University Press.
- Grady, Joseph & Oakley, Todd & Coulson, Seanna. 1999. Blending and metaphor. In Gibbs, Raymond W. Jr. & Steen, Gerard J. (eds.), *Metaphor in cognitive linguistics*, 101–124. Amsterdam: John Benjamins.
- Heywood, John & Semino, Elena & Short, Mick. 2002. Linguistic metaphor identification in two extracts from novels. *Language and Literature* 11(1). 35–54.
- Kövecses, Zoltán. 2002. *Metaphor: A practical introduction*. New York: Oxford University Press.
- Lakoff, George & Espenson, Jane & Schwartz, Alan. 1991. Master Metaphor List. Berkeley, CA: Cognitive Linguistics Group University of California, Berkeley. (Technical report).
- Lakoff, George & Johnson, Mark. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.
- Lakoff, George & Turner, Mark. 1989. *More than cool reason: A field guide to poetic metaphor*. Chicago: University of Chicago Press.
- Littlemore, Jeannette & Tagg, Caroline. 2016. Metonymy and text messaging: A framework for understanding creative types of metonymy. *Applied Linguistics* 39(4). 481–507.
- Maldonado, Concepción (ed.). 2012. *Diccionario CLAVE: Diccionario de uso del español actual* [CLAVE dictionary: Dictionary of Spanish contemporary usage]. Cataño: Ediciones SM.
(<http://clave.smdiccionarios.com/app.php>) (Accessed 2019-03-15).
- Markert, Katja & Nissim, Malvina. 2003. Corpus-based metonymy analysis. *Metaphor and Symbol* 18. 175–188.
- Ogarkova, Anna & Soriano, Cristina. 2014. Variation within universals: The ‘metaphorical profile’ approach to the study of ANGER concepts in English, Russian and Spanish. In Musolff, Andreas & MacArthur, Fiona & Pagani, Giulio (eds.), *Metaphor and intercultural communication*, 93–116. London: Bloomsbury.

- Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol* 22(1). 1–39.
- Radden, Günter & Kövecses, Zoltán. 1999. Towards a theory of metonymy. In Panther, Klaus-Uwe & Radden, Günter (eds.), *Metonymy in language and thought*, 17–59. Amsterdam: John Benjamins.
- Real Academia Española. 2001. *Diccionario de la lengua española* [Dictionary of the Spanish language]. (<http://www.rae.es/rae.html>) (Accessed 2019-03-15).
- 2009. *Nueva gramática española: Sintaxis II* [New Spanish grammar: Syntax II]. Madrid: Espasa Libros.
- Rundell, Michael & Fox, Gwyneth (eds.). 2002. *Macmillan English dictionary for advanced learners*. Oxford: Macmillan Education.
- Semino, Elena. 2008. *Metaphor in discourse*. Cambridge: Cambridge University Press.
- Semino, Elena & Demjen, Zsófia & Hardie, Andrew & Payne, Sheyla & Rayson, Paul. 2018. *Metaphor, cancer and the end of life: A corpus-based study*. London: Routledge.
- Semino, Elena & Heywood, John & Short, Mick. 2004. Methodological problems in the analysis of metaphors in corpus conversations about cancer. *Journal of Pragmatics* 36(7). 1271–1294.
- Soriano, Cristina. 2012. La metáfora conceptual [Conceptual metaphor]. In Ibarretxe-Antuñano, Iraide & Valenzuela, Javier (eds.), *Lingüística cognitiva*, 97–120. Barcelona: Anthropos.
- Steen, Gerard J. 1999. From linguistic to conceptual metaphor in five steps. In Gibbs, Raymond W. Jr. & Steen, Gerard J. (eds.), *Metaphor in cognitive linguistics*, 55–77. Amsterdam: John Benjamins.
- 2007. Finding metaphor in discourse: Pragglejaz and beyond. *Cultura, Lenguaje y Representación / Culture, Language and Representation* 5. 9–25.
- Steen, Gerard J. & Biernacka, Ewa & Dorts, Aletta G. & Kaal, Anna & López Rodríguez, Clara I. & Pasma, Trijntje. 2010a. Pragglejaz in practice: Finding metaphorically used words in natural discourse. In Low, Graham & Todd, Zazie & Deignan, Alice & Cameron, Lynne (eds.), *Researching and applying metaphor in the real world*, 165–184. Amsterdam: John Benjamins.
- Steen, Gerard J. & Dorts, Aletta G. & Herrmann, Berenike J. & Kaal, Anna & Krennmayr, Tina & Pasma, Trijntje. 2010b. *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam: John Benjamins.
- Stefanowitsch, Anatol. 2006. Words and their metaphors: A corpus-based approach. In Stefanowitsch, Anatol & Gries, Stefan Th. (eds.), *Corpus-based approaches to metaphor and metonymy*, 61–105. Berlin: Mouton de Gruyter.
- Valenzuela, Javier & Soriano, Cristina. 2005. Cognitive metaphor and empirical methods. *Barcelona English Language and Literature Studies (BELLS)* 14. 1–19.

Vv. Aa. 2005. *El libro de Radio Nikosia: Voces que hablan desde la locura* [The Radio Nikosia book: Voices speaking from madness]. Barcelona: Gedisa.

Contact information:

Marta Coll-Florit
Salvador Climent
Faculty of Arts and Humanities
Universitat Oberta de Catalunya
Av. Tibidabo 47
08035 Barcelona
Catalonia, Spain
e-mail: mcollfl(at)uoc(dot)edu
 scliment(at)uoc(dot)edu

Assessing the productivity of the Estonian deverbal suffix *-mine* in five registers of Estonian

Maarja-Liisa Pilvik
University of Tartu

Abstract

This article takes a usage-based quantitative approach to assess the morphological productivity of the deverbal action nominal suffix *-mine* (e. g. *magamine* ‘sleeping’) in Estonian newspaper texts, fiction, scientific texts, spoken regional dialects, and spoken spontaneous common language. While it is possible to derive an action noun with *-mine* from every verb, its productivity in different registers varies according to the aspects highlighted by the different measures: realized, expanding, and potential productivity. In addition to these measures, the ratio of the verb stems realized as verbs, and the stems realized as *-mine* nouns is compared to detect stems which are attracted to the nominalized structure more than would be expected by their overall occurrence in the corpora. The results of this study indicate that even derivation suffixes with ‘absolute’ productivity vary in terms of their contribution to the growth rate of the vocabulary of a given register.

Keywords: corpus linguistics, derivation, morphological productivity, nominalization, Estonian

1 Introduction

The suffix *-mine* (e. g. *kakle-mine* ‘fighting’) is considered to be the most regular and productive means for deverbal nominalization in Estonian (Erelt et al. 1993; Kasik 2009; Erelt 2017). Regularity is generally understood to refer to the formal, syntagmatic aspect of the derivation process, i. e. the fact that it is possible to derive a noun from every verb stem in exactly the same way using this suffix. *-mine* always attaches to the strong stem of the verb and therefore can be easily formed from the 1st infinitive form (*hüppa-ma* jump-1INF ‘to jump’ → *hüppa-mine* jump-NM.NOM ‘jump-ing’ vs. *hüpa-ta* jump-2INF ‘to jump’). While the lexical category changes, the semantics generally does not, and the relationship between the derivative and

the underlying verb is transparent. Productivity, in turn, is defined somewhat differently in morphological approaches to word formation, depending also on the status of the language as an object of study, and may also include the notion of regularity (e. g. Pinker 1999; in Estonian, Kasik 2009; 2013). According to Rainer (1987, cited in Gaeta & Ricca 2015), as many as six meanings can be associated with the notion of morphological productivity alone:

1. the number of words formed with a certain word-formation rule;
2. the number of novel words created with a certain word-formation rule;
3. the possibility of creating new words with a certain word-formation rule;
4. the probability of creating new words with a certain word-formation rule;
5. the number of possible words formed with a certain word-formation rule;
6. the relation between occurring and possible words formed with a certain word-formation rule.

Even more extensive lists have been made with regards to the meanings of ‘productive’ in linguistics in general (cf. Barðdal 2008: 10–11 for a list of 19 different usages). Regardless of the exact definition, the notion of morphological productivity is most often used when distinguishing between morphological categories with a more or less fixed membership and those with a growing membership. While the suffix *-mine* surely belongs to the latter class due to the regularity of the word-formation rule, its productivity would vary with regards to at least some of the 6 aforementioned definitions. The different units of assessment, such as *number*, *probability*, and even the abstract *possibility*, imply that the productivity of a category can be understood as either something fixed, relational, or purely potential/theoretical.

In this article, I follow the general understanding in usage-based linguistics, that productivity is the probability of a grammatical pattern being used to create novel structures (e. g. Bybee 2001; Baayen 2003). It has been shown that a productive morphological pattern is characterized by a large number of *hapax legomena*, i. e. lexemes/types in the category occurring only once in a given corpus since the productivity of the word-formation rule also guarantees the understanding of new entities. The distribution of lexemes of a non-productive category, however, is completely different: there are many high-frequency

lexemes, fewer rare and very few (if any) lexemes which occur only once (Baayen & Renouf 1996; Baayen 2009). The notions of *type* and *hapax legomena* are utilized also in more precise measures developed to highlight different aspects of productivity in quantitative approaches to morphological productivity (e. g. Baayen 1992; 1993; Baayen & Renouf 1996).

It has also been demonstrated that in addition to a number of syntactic and lexical phenomena (Biber 1995), also certain derivative affixes may be more characteristic of certain types of texts, although as Plag et al. (1999) note, nominalization does not appear to play a crucial role in register differentiation. Nevertheless, it has been noted that in Estonian, derivatives with the suffix *-mine* are especially characteristic of written scientific, legal, and administrative texts (Kerge 2003), but have spread also to journalistic texts (Kasik 2006).

Although the derivational patterns (incl. *-mine* derivation) and their productivity in Estonian have been extensively described (e. g. Vare 1994; Kasik 2009), frequencies and their distributions in different registers have not been empirically and systematically accounted for.¹ This is probably because productivity has been associated with language capacity rather than language use (similarly to e. g. Anshen & Aronoff 1999 and Dressler 2003), and the role of frequency has only been seen in the context of lexicalization or semantic specialization of derivatives (Kasik 2011). However, by defining productivity through the concept of probability, the productivity of a derivative pattern can be considered scalarized, and usage frequencies could be operationalized not only to identify lexicalized linguistic units, but also to study the variation in the productivity of certain derivational patterns across different types of texts.

The purpose of the current study is to approach morphological productivity from a quantitative, usage-based perspective and to evaluate the productivity of the suffix *-mine* in different Estonian sub-languages and registers represented in the Balanced Corpus of Estonian (written newspaper texts, fiction, and scientific texts), the Corpus of Estonian Dialects (oral traditional dialects), and the Phonetic Corpus of Estonian Spontaneous Speech (oral common language). To examine the differences between the five registers, I use the realized, expanded, and potential productivity (e. g. Baayen

¹ Krista Kerge's studies (2002; 2003), where the quantitative analysis of the complexity of text types implicitly also included productivity as the ratio between token count of *-mine* nouns and the number of tokens in the corpus, may be considered as exceptions here. Kerge also studied the change of this ratio diachronically, providing an interesting insight into the effect of language policies and ideological orientation on the use of *-mine* nouns.

1992; 2009; Plag et al. 1999) of *-mine* derivational pattern. Based on e. g. Plag et al. (1999: 218), I hypothesize that derivation, which is said to contribute the most to the growth of vocabulary, is much less productive in spoken registers than in written registers. I will also study the ratios of verbal stems realized as verbs and *-mine* nominalizations in the different corpora to unveil the (dis)preferences of the suffix *-mine* with regards to the stems to which it attaches.

In what follows, I will first describe *-mine* nominalization in the light of its formation and main functions. Then, I will discuss the aforementioned measures that have been proposed for assessing morphological productivity. This will lead to presenting my data and comparative findings for the Estonian suffix *-mine*. Finally, I will discuss the implications of these findings to the general understanding of the productivity of *-mine* in Estonian and the pitfalls of the quantitative corpus-based approach taken in this article.

2 Form and functions of *-mine* nominalization

The suffix *-mine* is the most common means for deverbal predicate nominalization in Estonian (Erelt et al 1993; Erelt 2017) and translates roughly as *-ing* in English. Syntactically, it is used to create a secondary predication for using a clause as an argument or modifier in another clause. The agent (1) or the patient (2) of the underlying predicate is usually expressed as a genitive modifier in the resulting noun phrase. The patient can also form a compound with the action nominal (3), in which case the patient specifies a type of action expressed by the verb.

- (1) *isa peksa-b* → *isa peks-mine*
 father beat-3SG father.GEN beat-NM.NOM
 ‘father is beating’ ‘father’s beating’
- (2) *isa peks-t-akse* → *isa peks-mine*
 father.PART beat-IPS-PRS father.GEN beat-NM.NOM
 ‘father is being beaten’ ‘the beating of the father’
- (3) *isa+peks-mine*
 father+beat-NM.NOM
 ‘fatherbeating’

Formally, the resulting forms are considered to belong to the noun word class, since they have a full number and case paradigm, they can take genitive, pronominal or adjectival attributes (including those not in the structure of the underlying predicate), and among other things, can function syntactically also as subjects, objects and predicatives of the clause. These NPs also exhibit a number of verbal properties, such as being able to retain the underlying verb's adverbial modifiers and predicatives or sometimes even govern the object (cf. Pilvik 2016; 2017). Due to this categorial ambiguity, the “productively” formed action nominals are sometimes distinguished from the lexicalized, aspectually bounded, or semantically specialized verbal nouns (Kasik 2009; Erelt 2017: 831–832). Indeed, *-mine* nominalizations do not always refer to the aspectually unbounded processes, but also to single events as well as the results or objects of these events (Kasik 2009). Those structures have been said to be idiomatized, i. e. semantically opaque, and do not strictly fit in the action nominal category (Erelt 2017: 832). The properties which are usually linked to the lexicalized meanings are pluralization (*vali-mise-d* ‘elections’ vs. ‘electings’), absence of the predicate's obligatory arguments (*tead-mine* ‘knowledge’ vs. ‘knowing’), and incorporating a specifier (*söögi-tege-mine* ‘cooking’ vs. ‘foodmaking’) (Kerge 2001: 42–43). However, the mere presence of one or several of these attributes does not suffice to distinguish between the processual and referential *-mine* nouns, nor is there enough reason for differentiating between the suffix producing processual nominalizations and the one responsible for just the idiomatic structures. Instead, defining processuality as a gradient and highly context-related property of *-mine* nouns seems more appropriate to account for all the possible aspectual and referential readings in addition to the strict reading of action (Kerge 2001: 40; Pilvik 2017: 304). Also, while productivity characterizes the whole derivational pattern, idiomatization only concerns specific derivatives, which can acquire additional semantic properties in addition to their categorial meaning (Kasik 2009). This implies that the more lexicalized units could be detected based on their type frequencies as it is well-known that opaque forms tend to be high-frequency words.

Research on *-mine* nominalization in Estonian seems to fall largely into three: studies which focus on the syntactic function of the nominalization (i. e. applying the nominalization suffix to use a clause in the position of a noun, Kasik 1968; 1975; Kerge 2001), studies which focus on its lexical function (i. e. applying the nominalization suffix to create new terms, Kerge 2002; 2003), and studies covering its textual functions such as abstracting,

generalizing, or thickening the text (Kasik 1995; 2006). All functions seem to operationalize a productive rule, but they do that for different purposes. The functions for which *-mine* nouns are created have been shown to depend on the register. As most studies have been done based on written language, the functions considered have been the ones prevailing in namely written domains of language use. Thus, legal texts mostly make use of *-mine* nouns for the syntactic purpose as well as for thickening the text (Kerge 2002: 38); same functions apply in journalistic texts (Kasik 2006), although the need for abstract concepts is also mentioned (Kerge 2002); academic and scientific texts mostly use *-mine* nouns for the need of specific terminology (Kerge 2002; 2003). Not much has been said about spoken Estonian, but one can assume that the very conditions under which spoken and written language are produced also affect the functions in which nominalization is used. A study on dialect data, for example, demonstrated the discourse-related functions of *-mine*, such as anaphoric referencing, relaxing the processing load, or providing time for utterance planning (cf. Pilvik 2017).

In addition to functions, different registers also tend to be used to communicate different kinds of topics (Baayen 2009), which affects the distribution of lexemes affected by suffixation. The effects of the functional and topical bias have to be taken into account when interpreting productivity measures described in the next subsection.

3 Measuring productivity

In a probabilistic approach to morphology, productivity is not affected only by syntagmatic regularity, but also by schemas, constructions, and generalizations. According to Baayen (2009: 901), “morphological productivity can be understood as resulting from a great many factors such as the individual language user’s experience with the words of her language, her phenomenal memory capacities, her conversational skills, her command of the stylistic registers available in her language community, her knowledge of other languages, her communicative needs, her personal language habits and those of the people with which she interacts”.

Several corpus-based measures have been developed in order to seek evidence for intuitions about morphological productivity of certain patterns, catch different aspects of productivity, and to provide comparable statistics across multiple genres and registers. I exploit the three measures discussed in

e. g. Baayen (1993), Baayen & Renouf (1996), and Baayen (2009) – realized productivity, expanding productivity, and potential productivity – since these have been extensively used in corpus studies concerning the morphological productivity of affixes in a wide range of languages. Corpus-based quantitative studies on derivational productivity have focused mainly on the differences between several derivational affixes in Indo-European languages, such as English, Dutch, and Italian, both within one register and across multiple registers (e. g. Baayen 1993; Plag 1999; Plag et al. 1999; Gaeta & Ricca 2006). Using these measures to study the variation in productivity for one suffix across registers has been less common. For Finnic languages, the measures have been applied in e. g. experimental studies in Finnish morphology (Mäkisalo 2000; Järvikivi 2003; Nikolaev & Niemi 2008).

Realized productivity C of a pattern refers to *the contribution* of a category (e. g. *-mine* derivatives) to the size of the whole vocabulary. It is also known as the *extent of use* (Baayen 1993) and is estimated by the type count (V) of the category. Since the productivity measures are compared across registers with slightly differing token counts in this article, *realized productivity C* here refers to the quotient of the number of types (V) in a given category and the total number of tokens (N) in a corpus. Realized productivity is therefore a relational measure in this article, instead of an absolute count. The higher the value, the more a given pattern occurs in the vocabulary and the more productive it can be considered.

*Expanding productivity P** of a pattern indicates *the expansion rate* of a given category. This is also known as the *hapax-conditioned degree of productivity* and is calculated by dividing the number of hapax legomena in a given category ($n1$) by the number of hapaxes in the whole corpus ($N1$). The higher the value, the more a given category attracts new structures (compared to other possible categories) and the more productive it can be considered. In probabilistic terms, expanding productivity P^* shows the conditional probability that a word belongs to a certain category, given that this word occurs in the corpus only once.

Finally, *potential productivity P* reflects *the growth rate of the vocabulary* in a given category. It is also called the *category-conditioned degree of productivity* and is calculated by dividing the number of hapax legomena in a given category ($n1$) by the number of tokens in this category (Nc). The higher the value, the more lexically diverse the category is and the more productive it can be considered. In probabilistic terms, potential productivity shows the conditional probability, that a word occurs in the corpus only once, given

Table 1. Three measures of morphological productivity

Measure	Calculation	Reflects
Realized productivity	$C = V/N$	Contribution of the category to the size of all vocabulary.
Expanding productivity	$P* = nI/NI$	Attraction of novel structures by the category.
Potential productivity	$P = nI/Nc$	The potential of creating new structures within the category.

that it belongs to a certain category. This measure is designed to reflect the intuition that it is easier to think of neologisms with one suffix (e. g. *-mine*) than it is with another (e. g. another Estonian deverbal suffix *-us*) (Baayen 2009). Table 1 presents the summary of the three measures.

It is well-known that words with less compositional meaning tend to have higher usage frequencies, but a binary division between unproductive lexicalized elements (which are stored in lexicon) and productive schemas (which are stored in grammar) is usually not appropriate, especially with *-mine* nominalization which produces structures on the whole scale from non-referential to fully referential (cf. Pilvik 2017). It has also been shown with experiments that all complex words (not only irregular or "lexicalized" words) leave traces in lexical memory (Hay & Baayen 2005). In order to compare the likelihood of encountering nouns with a more or less fixed meaning in different registers, I have calculated what I call here *NV-score* for each individual verb stem in the samples of the five different subcorpora. NV-score of a verb stem is the difference between the proportion of that stem among the *-mine* nouns and the proportion of that stem among verbs. When this difference is 0, the likelihood of this stem occurring in the corpus as a verb is equal to the likelihood of it occurring as a *-mine* noun. When the difference is > 0 , the stem is less likely to occur in a nominalized structure than would be expected given its probability to occur as a verb in the corpus. When the difference is < 0 , the stem is more likely to occur in a nominalized structure than would be expected. The reason for doing this instead of simply looking for the most frequent types in each corpus is to eliminate the effect of the above-mentioned topical bias in different registers. The type-based NV-score provides additional perspective to straightforward productivity measures: if

some verbs are more likely to be nominalized with *-mine* than would be expected by their overall frequency in the corpus, then we can assume the corresponding nominalized structures are possibly also more entrenched in the lexicon. The admittedly simple reasoning behind this assumption relies on the regularity of the word-formation pattern: if all verbs can be nominalized with *-mine*, then *-mine* nouns created for syntactic functions should show similar distribution to verbs with regards to the lexical stems from which they are formed. In turn, *-mine* nouns created for lexical or textual purposes are independent of the verbs used to discuss certain topics in a given register. The stems rarely nominalized are expected to exhibit semantic properties, which make them less likely to be construed as an entitized referent instead of a situational one, or belong to the class of verbs often used in non-normal clauses such as existential, possessive, experiencer, or resultative clause.² Formally, however, no lexical restrictions exist for *-mine* nominalization.

In recent decades, a wealth of researchers have turned from corpus data to experimental studies in order to provide even stronger support for the gradual and probabilistic view of morphology (cf. an overview in Hay & Baayen 2005). In these approaches, the morpheme as a theoretical construct is backgrounded, and instead, full words are viewed as the basic units in the lexicon. It is paradigmatic relations, instead of a set of lexical entries and rules, which help characterize lexical representations and determine lexical processing. In the context of *-mine* nouns, this would mean accounting for the support a derived noun gets from other words occurring in the same morphological family (*laul-mine* sing-NM ‘singing’, *laul-ma* sing-1NF ‘to sing’, *laul-da* sing-2INF ‘to sing’, *laul-ja* sing-AN ‘singer’ etc.) to explain e. g. reaction times and precision in lexical decision or word naming tasks. This highly intriguing perspective remains outside the scope of this article, which at this stage limits itself only to corpus data. However, considering the frequency of individual verb stems (covering all the verbs’ inflectional forms) in different corpora is also a step towards higher paradigmatic awareness when assessing the degree of lexicalization.

² Erelt et al. (1993: 269) make the claim that verbs which are never used in Estonian normal clauses (e. g. *piisama* ‘to suffice, to be enough’ in *Pildi tegemiseks piisab ka tavalisest kaamerast* ‘An ordinary camera is enough to take a photo’) can not be nominalized. While I do not hold this to be true, I do expect non-normal clauses to be nominalized less frequently.

4 Data

The data for this study comes from 3 different corpora, representing altogether five different registers of Estonian. As the aim of this article is to provide a global view of the variation in probabilistic productivity of the suffix *-mine*, the subdivision into registers is taken to be predefined by the corpora and a more detailed view of the different text types (e.g. sports commentaries against interviews) is currently ignored.

The Balanced Corpus of Estonian³ (BCE) comprises 5 mln tokens from newspapers, 5 mln tokens from fiction, and 5 mln tokens from scientific texts. The newspapers subcorpus (NEWS) dates from 1995 to 2007 and contains both daily and weekly newspapers; fiction subcorpus (FICT) contains excerpts from Estonian literature, both prose and poetry, from the period between 1987 and 2011; the scientific subcorpus (SCI) holds Ph. D. dissertations and articles from scientific journals, covering a wide range of disciplines, and dates from 1995 to 2006. BCE is a subpart of the big Mixed Corpus of Estonian and is compiled first and foremost to enable the comparison of the genres and text classes of written Estonian. The corpus also has a morphologically annotated version, where each token has received automatic morphological analysis.

The Corpus of Estonian Dialects⁴ (CED) comprises about 900,000 tokens of manually morphologically annotated transcriptions of recordings from the 1960s–1970s, covering all 10 Estonian traditional dialect areas (DIA). The recordings are unstructured interviews, where elderly informants speak about their everyday life, childhood, customs, traditions, past events, and work. The transcripts follow the conventions of the dialect corpus's simplified transcription.⁵ However, as both the original transcriptions (following Finno-Ugric transcription rules) as well as morphological annotation have been done manually and neither strictly follows the orthographic rules of written Estonian, there are also some idiosyncracies. For example, compound words are written using both separate writing (*ära viimine* 'taking away') and solid writing with the transcription sign + (*ära+viimine* 'taking away').

The Phonetic Corpus of Estonian Spontaneous Speech⁶ (PCESS) is the

³ <https://www.cl.ut.ee/korpused/grammatikakorpus/> (accessed 2018-01-01).

⁴ <https://www.keel.ut.ee/et/keelekogud/murdekorpus/> (accessed 2015-09-29).

⁵ In addition to phoneme realizations, simplified transcription also marks short and long pauses, liaisons, quantity alternations and compound words; https://www.keel.ut.ee/sites/default/files/www_ut/emk_teejuht2015.pdf (accessed 2019-03-01).

⁶ <https://www.keel.ut.ee/en/languages-resourceslanguages-resources/>

Table 2. Token counts, type/token ratios, and growth rates of the (sub)corpora⁷

(Sub)corpus	Corpus Size in tokens	Type/token ratio (TTR)	Growth rate
NEWS	4,675,823	0.045	0.024
FICT	4,953,823	0.033	0.018
SCI	4,798,966	0.048	0.025
DIA	890,788	0.028	0.013
SP	338,807	0.052	0.027

smallest corpus in this study, comprising only around 340,000 tokens (SP). The corpus contains 60 hours of speech recordings, both dialogues between familiar people and monologues, recorded between the years 2006 and 2016. Most recordings are done in a studio, some have also been done during fieldwork. The average length of the recordings is 30 minutes. As the name of the corpus suggests, the purpose of its compilation was to provide data on the phonetic traits of spontaneous speech (e. g. speech sounds, sound structures, syllables, assimilation, voice quality), but its accessibility and the abundance of different segmentation layers (incl. morphological annotation) makes this corpus a valuable resource also for non-phoneticians.

Table 2 shows the size of the corpora in morphologically annotated tokens⁸ and the type/token ratios (TTR), which are considered a good indication of the range of vocabulary used in the corpus under consideration. A high TTR indicates more lexical variation and a low TTR relatively little lexical variation (Baker et al. 2006: 162). The growth rate in the last column is calculated by dividing the number of hapaxes by the total number of tokens and translates as the probability of the occurrence of a type not encountered before in the corpus.

Simple type/token ratio (TTR) is very sensitive to the size of the text or

phonetic-corpus-estonian-spontaneous-speech (accessed 2018-09-02).

⁷ Types are counted according to their specified part-of-speech to decrease the chance that homonymic forms, which represent different linguistic categories, will be counted as one type. While *palk* ‘log’ and *palk* ‘payment’ will still remain one type, *või* ‘or’ and *või* ‘butter’ will not. This, however, can artificially increase the type count where manual annotation has been involved.

⁸ Punctuation is excluded.

Table 3. Token counts, type/token ratios, and growth rates of subcorpora

Sampled corpus	Sample size in tokens	Type/token ratio (TTR)	Growth rate
NEWS	336,984	0.123	0.068
FICT	398,796	0.075	0.041
SCI	337,696	0.100	0.052
DIA	335,325	0.041	0.019
SP	338,807	0.052	0.027

corpus. The longer the text, the more types have already been encountered and the lower the likelihood of any given token representing a new type. Therefore, it is crucial to provide texts of equal sizes when comparing their TTR (Hardie & McEnery 2006: 139). Productivity measures described above also rely on type frequencies. To enable comparison between different registers, I took a random sample from each corpus, with reference to the size of the smallest corpus, the PCESS.⁹ The sampling of the files in each corpus stopped when the token count in that sample reached higher than 335,000 (condition was checked prior to sampling another file). While these samples can be considered rather small, they are still able to shed some light on the variation of productivity between the registers.¹⁰ The sample sizes, their TTRs and growth rates are presented in Table 3.

As the comparison of Table 2 and Table 3 reveals, the TTRs and growth rates are indeed considerably higher in the samples than in the full-sized corpora (except for the spoken spontaneous language, where sample size equals corpus size) and sampling, therefore, is necessary for adequate comparisons. Table 3 shows that both TTR and growth rate are higher in the written registers than in the spoken registers. This means that written language

⁹ A similar approach, although a more refined one, was taken in Plag et al. (1999), where sub-corpora were compared for the largest range of token sizes that they had in common. They also took into account the shape of vocabulary growth curve and assessed the number of types and productivity measures by calculating the average of each measure at twenty equally spaced intervals. In this article, however, the aspects of productivity are assessed only at the final values of the sample sizes.

¹⁰ I rely on e. g. Plag et al. (1999: 215), who have written that “What exactly counts as sufficiently large is not easy to determine, but even relatively small corpora like the Dutch Eindhoven Corpus (600,000 words of written text) seem to yield interesting results”.

corpora are lexically more diverse and also expanding their vocabulary at a higher rate.

From the samples, I exhaustively extracted all occurrences of noun lemmas formed with the suffix *-mine*. Extracting *-mine* nouns was a seemingly effortless task, since there are no morphological alternations (*valmista-ma* ‘to make’ → *valmista-mine* ‘making’, *arutle-ma* ‘to discuss’ → *arutle-mine* ‘discussing’, *uuri-ma* ‘to study’ → *uuri-mine* ‘studying’, *hari-ma* ‘to educate’ → *hari-mine* ‘educating’). The morphologically annotated corpora, thus, enabled to simply look for lemmas which ended with *-mine* and were marked for being nouns.¹¹ The lists were manually cleaned from foreign words (e. g. *examine*, *phenylpropanolamine*). The extraction of deverbial *-us* nouns, for example, would be somewhat more difficult, because *-us* can also be used to derive nouns from adjectives (e. g. *haige* ‘ill, sick’ → *haig-us* ‘illness, sickness’) and the derivation is subject to both stem and suffix alternations (e. g. *valmista-ma* → *valmist-us* ‘making’, *arutle-ma* → *arutl-us* ‘discussion’, *uuri-ma* → *uuri-mus* ‘(a) study’, *hari-ma* → *hari-dus* ‘education’). However, going through the word lists and manually correcting the mistakes also for the suffix *-mine* was crucial for the measures to be informative. Especially since with rare words, the automatic morphological analyzer does not (yet) have the capacity to always understand these structures the same way that a human interpreter would. Therefore, it occasionally proposes analyses which are not appropriate and skew the statistics.

Finally, I collected the stems of all verbs in the samples in order to see, whether the distribution of the nominalized verb lemmas corresponds to the overall distribution of the verbs used in certain registers.

5 Results

Corpus data allow examining, how the morphological productivity of the suffix *-mine* varies across registers. Each of the three measures discussed provides a slightly different perspective on this variation. As the magnitudes of the measures are essentially proportions and can be interpreted as percentages, they vary according to the frequency of the phenomenon in question. Individual derivational affixes form rather specific grammatical categories and cannot, therefore, account for a very large proportion of the

¹¹ The suffix *-mine* can also non-productively derive adjectives (e. g. *alumine* ‘lower; lowermost’, *pealmine* ‘upper; uppermost’).

Table 4. Realized productivity for the suffix *-mine* in 5 registers

Register	Realized productivity
NEWS	0.0031
FICT	0.0019
SCI	0.0030
SP	0.0012
DIA	0.0007

overall words. Additionally, the NV-score will help detect the verbs which are considerably more or less likely to be nominalized than would be expected by their average presence in the corpora.

The productivity measures are calculated based on the samples of approximately 335,000 tokens per corpus, which is very small, compared to the corpora used to study e. g. English derivational suffixes. However, as productivity in spoken registers has so far largely been neglected in Estonian linguistics, the results based on even relatively small datasets are extremely valuable.

5.1 Realized productivity of *-mine*

In this article, realized productivity means finding the ratio between the number of types of *-mine* nouns and the total number of tokens in the corpus. The measure reflects the contribution of the *-mine* nominalization category to the size of the whole vocabulary in the corpus, or in other words, the productivity of a morphological process in the past. The values for realized productivity are presented in Table 4.

The lowest realized productivity for the suffix *-mine* is in the dialect corpus (0.0007), the highest in news and scientific writing. This means that out of the five registers, *-mine* nouns as a grammatical category contribute the most to the size of vocabulary in these written text types where *-mine* nouns most likely function as means for abstraction and generalization, thickening the text, or providing new terms. The formation rule seems to be the least productive in dialect data with respect to the category's contribution as a whole. These results are in accordance with what is generally claimed for derivational patterns: they are much less productive in spoken

registers than in written ones. However, it seems that written fiction is somewhat closer to spoken common language. Both registers are used to communicate less specific purposes than newspapers or scientific texts. Therefore, realized productivity is a measure which depicts the different functions of *-mine* derivation and is closely linked to the written-oral dimension of communication. The differences between the registers seem quite small, even when considering that *-mine* derivation is not the most frequent phenomenon among morphological categories. When translated into percentages, the difference between the register with the highest realized productivity (NEWS) and one with the lowest (DIA) is only 0.24 percentage points. However, the absolute type frequencies of 1060 and 232, respectively, do speak of a significant difference.

5.2 Expanding productivity of *-mine*

Expanding productivity or the *hapax-conditioned degree of productivity* is found when dividing the number of hapax legomena among *-mine* derivatives by the number of hapaxes in the whole corpus. The measure shows the rate with which a category gives rise to novel structures and is a more refined one compared to the previous measure. The values for realized productivity are presented in Table 5.

Expanding productivity shows the highest value again in a written register, but this time, it is fiction which exhibits high productivity of *-mine*, similarly to spoken common language, whereas scientific writing fell to the bottom of the list. A high expanding productivity means that if we were to add one extra token into our corpus, expressing some new concept, a *-mine* noun would be a plausible grammatical category for realizing that new structure. Consequently, this category will often require lexical procedural knowledge in both comprehension and production. Differences between registers imply that the relative contribution of *-mine* to the growth rate of the vocabulary in both written and spoken common language (FICT and SP, respectively) is an increasing function of the sample size with a far greater slope than, for example, in scientific writing. This is somewhat counterintuitive at first sight, since the terminological function of *-mine* in scientific texts should make it highly attractive to novel concepts. However, expanding productivity is a hapax-based measure. One can assume that terms in scientific writing are usually not created for single use, but are meant for denoting a specialized concept which is important to the subject of the text. Therefore, finding

Table 5. Expanding productivity for the suffix *-mine* in 5 registers

Register	Expanding Productivity
NEWS	0.0202
FICT	0.0274
SCI	0.0186
SP	0.0236
DIA	0.0204

specialized terms among hapaxes in scientific writing is less likely.

5.3 Potential productivity of *-mine*

The third and final measure, potential productivity, shows the growth rate of the vocabulary among *-mine* nouns and is found by dividing the number of hapaxes found with *-mine* by the total number of tokens with *-mine*. It reflects the likelihood of forming more types than are actually attested in the corpus. Potential productivity values for the five subcorpora are presented in Table 6.

The potential productivity measures show the most extensive variability amongst the three measures. Here, dialects exhibit the highest productivity for *-mine* and the differences between the registers are slightly clearer than for the other measures. In dialects, a considerable proportion (19%) of *-mine* nouns occur only once, indicating that the potential to encounter a *-mine* noun not attested before in the corpus is over nine times higher in dialects than it is in scientific texts and nearly three times higher than in newspapers. Fiction, once again, leans towards the spoken registers in terms of the more productive use of the suffix *-mine* than other written registers, making it a likely intermediate category between written and spoken genres. This measure is strongly affected by lexical diversity. As the dialect corpus contains data from 10 traditional dialect areas, the large proportion of hapaxes reflects lexical variability between these areas. A high number of types coined only once in spoken common language and fiction, in turn, reflects creativity in language use (slang, intentional wordplay etc.). Another interpretation of the measure involves stating the higher likelihood of using lexical procedural knowledge to formulate a preverbal concept (e. g. an action of some sort) in spoken dialect interviews than in newspaper texts or scientific writing. As this

Table 6. Potential productivity for the suffix *-mine* in 5 registers

Register	Potential productivity
NEWS	0.0764
FICT	0.1746
SCI	0.0260
SP	0.1358
DIA	0.1934

measure is restricted by one category, it is also more restricted as a measure for productivity with not being able to account for the potential for coining new structures with other means than the suffix *-mine*. Potential productivity is also said to be negatively affected by the amount of opaque words in a category (Baayen 2009), because of their tendency to be high-frequency words. Therefore, the low value for potential productivity in newspapers and scientific texts can be considered an indication of the presence of more *-mine* nouns with specialized meanings. Figure 1 summarizes the magnitude of different productivity measures in the five registers.

When distributed across the same scale, it can be observed that compared to potential productivity, the other two productivity measures differ less in magnitude across the five registers. Therefore, in comparing the productivity of *-mine* in these registers, the extent to which the regularity of the derivation pattern can be used to create *-mine* nouns in lexically diverse settings (irrespective of other available categories) seems to make the most difference. However, as the other two measures are relative frequencies based on considerably larger denominators, this bears no statistical significance.

5.4 NV-scores for the verbal stems

NV-scores reflect the degree to which certain verbal stems are attracted to the nominalized structure, given their overall frequencies in the different subcorpora. The overall distribution of the scores in the corpora is presented in Figure 2.

Figure 2 shows that most stems receive a score close to 0, while few stems rank considerably higher or lower. This is in accordance with the overall distribution of words in corpora, meaning that there are a lot of low-frequency

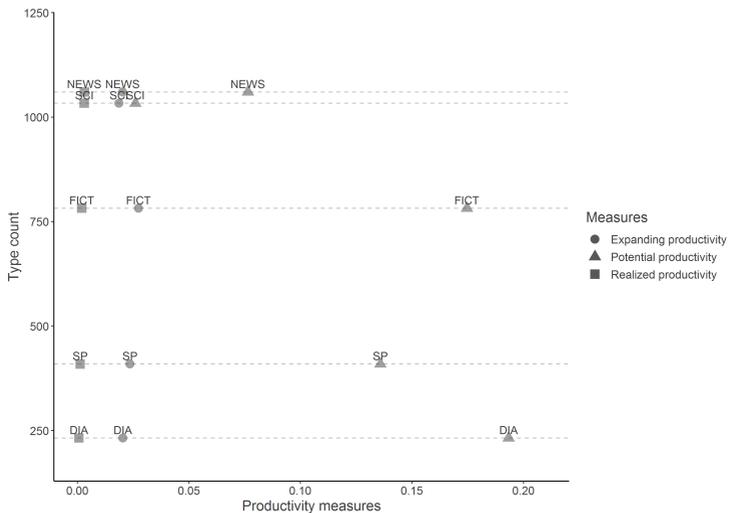


Figure 1. Comparison of three different productivity measures for the suffix *-mine*

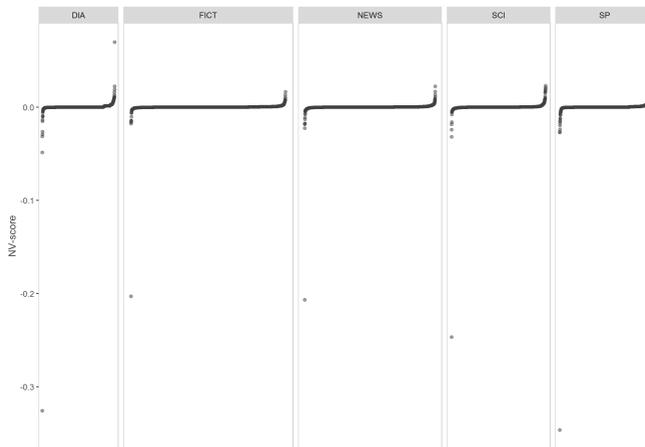


Figure 2. Distribution of NV-scores in the five registers

Table 7. Verb stems with highest NV-scores

NEWS	Score	FICT	Score	SCI	Score	DIA	Score	SP	Score
<i>kohtu-</i> 'meet'	0.022	<i>tege-</i> 'do'	0.017	<i>käitu-</i> 'behave'	0.023	<i>tege-</i> 'do'	0.070	<i>tege-</i> 'do'	0.057
<i>vali-</i> 'elect'	0.017	<i>liiku-</i> 'move'	0.013	<i>tead-</i> 'know'	0.021	<i>söö-</i> 'eat'	0.022	<i>liiku-</i> 'move'	0.034
<i>tege-</i> 'do'	0.013	<i>tead-</i> 'know'	0.011	<i>aruta-</i> 'discuss'	0.020	<i>ela-</i> 'live'	0.019	<i>luge-</i> 'read'	0.033
<i>kasuta-</i> 'use'	0.011	<i>ela-</i> 'live'	0.009	<i>kasuta-</i> 'use'	0.018	<i>aja-</i> 'drive; handle smth; cause smth'	0.016	<i>pakku-</i> 'offer'	0.017
<i>pakku-</i> 'offer'	0.010	<i>kohtu-</i> 'meet'	0.008	<i>kohalda-</i> 'apply; customize'	0.017	<i>peks-</i> 'beat'	0.013	<i>varieeru-</i> 'vary'	0.016

words, whose frequency divided by a larger number approaches 0, and few high-frequency words, which give slightly higher scores. The width of each facet reflects the overall number of different stems in the corresponding subcorpus. The five stems with the highest scores in each corpus are presented in Table 7.

The stems which receive the highest NV-scores and are thus strongly attracted to the nominalized structure tend to occur in more specialized meanings and constructions. For example, *kohtumine* 'meeting', *valimine* 'election', *liikumine* 'movement', *teadmine* 'knowledge', *pakkumine* 'offer' are more frequently used as referring to events, or the objects or results of the process, rather than to the processes themselves. A closer examination of the observations in the data reveals that these are the nouns also more actively used in the plural and without the predicate's arguments, fulfilling at least two of the criteria for higher degree of lexicalization. The curious case of *tege*- 'do', in turn, can be explained by its use in special constructions. In written language, *tegemine* is often used in the predicative construction *olema* 'be' + *tege*-NM_{PART} + PRED_{COM} (examples 4 and 5).

- (4) *Tege-mis-t ol-i nalja-ga.*
 do-NM-PART be-IPF.3SG joke-COM
 'It was a joke.'

- (5) *Tema näol on tege-mis-t varga-ga.*
 she.GEN POSTP be.3SG do-NM-PART thief-COM
 ‘She is a thief.’

In spoken registers, however, this construction is rare and instead, *tege-* occurs in the so-called *busy-construction* $A_{ADE} + olema$ ‘be’ + *tege-NM_{PART}* (+ X_{COM}), illustrated by example (6), where the agent expressed as an adessive experiencer is doing something keeping her busy. The optional argument in comitative can be interpreted as the patient affected by the agent’s activity. This construction is probably specialized from a general modal construction $A_{ADE} + olema$ ‘be’ + $V-NM_{NOM/PART}$ in example (7) (cf. Pilvik 2016).

- (6) *Ema-l on (looma-de-ga) tege-mis-t.*
 mother-ADE be.3SG animal-PL-COM do-NM-PART
 ‘Mother is busy (with the animals)’
- (7) *Anne-l ol-i kõvasti õppi-mis-t, et teis-te-le järele jõu-da.*
 Anne-ADE be-IPF.3SG hard study-NM-PART that other-PL-ADE ADV
 catch_up-2INF
 ‘Anne had to study hard to catch up with the others.’

The productivity of *-mine* nouns in different syntactic constructions (cf. Neetar 1988; Sahkai 2011; Pilvik 2016; 2017) as well as the productivity of the constructions themselves definitely calls for a closer inspection. However, this additional perspective does not fit in the scope of this article.

The five stems less likely to occur in *-mine* nominalizations are presented in Table 8. The table contains stems which occur proportionally considerably more as verbs than they do as *-mine* nominalizations. Not surprisingly, these are the most frequent verb stems which are multifunctional, often used as grammatical verbs and semantically generic, but also the verbs typically used in non-normal clauses (cf. § 3). While all of these stems do also occur in nominalized structures, where *-mine* nouns are processual and most likely used for syntactic purposes, their overall high frequency as verbs keeps them at the low end of the NV-scores. An exception here is the modal verb stem *või-* ‘can; may’, since modal verbs are unlikely candidates for nominalization when they don’t also carry a non-modal meaning (e. g. *saa-* ‘get; can’).

Table 8. Verb stems with lowest NV-scores

NEWS	Score	FICT	Score	SCI	Score	DIA	Score	SP	Score
<i>ole-</i> 'be; have'	-0.207	<i>ole-</i> 'be; have'	-0.203	<i>ole-</i> 'be; have'	-0.247	<i>ole-</i> 'be; have'	-0.326	<i>ole-</i> 'be; have'	-0.346
<i>saa-</i> 'get; can'	-0.022	<i>tule-</i> 'come'	-0.018	<i>või-</i> 'can; may'	-0.032	<i>saa-</i> 'get; can'	-0.048	<i>mine-</i> 'go'	-0.027
<i>tule-</i> 'come'	-0.018	<i>mine-</i> 'go'	-0.016	<i>pida-</i> 'must; keep'	-0.024	<i>mine-</i> 'go'	-0.031	<i>tule-</i> 'come'	-0.026
<i>või-</i> 'can; may'	-0.018	<i>saa-</i> 'get; can'	-0.015	<i>saa-</i> 'get; can'	-0.018	<i>tule-</i> 'come'	-0.030	<i>saa-</i> 'get; can'	-0.024
<i>ütle-</i> 'say'	-0.017	<i>või-</i> 'can'	-0.015	<i>tule-</i> 'come'	-0.016	<i>pane-</i> 'put'	-0.026	<i>ütle-</i> 'say'	-0.020

6 Discussion

The aim of this study is to examine the productivity of the suffix *-mine* in different registers of Estonian and use available corpus data in order to provide concrete measures gauging different aspects of productivity. This means going further than simply presenting token frequencies or posing intuitive statements based on the number of grammatical restrictions for a given morphological pattern. Accounting for the necessity of equally-sized corpora for comparisons as well as looking into the lexical distribution of the corpora allows making more targeted and specific claims about both the functional as well as structural characteristics of each register.

Findings from this study clearly show that the three measures of productivity used enable describing the productivity of *-mine* in more detail and stress important aspects of register-specific variation. Out of the three measures, the realized productivity meant for gauging the contribution of the *-mine* category to the size of the whole vocabulary showed the least differences between the five registers. The main distinction emerged between the written and oral registers: *-mine* in the three written registers exhibited higher realized productivity than in the two oral ones. This measure could also most clearly be linked to the traditional functions of *-mine* nouns in different registers: high productivity in newspapers and scientific writing can be attributed to the necessity to use nominalization as means for abstraction and

creating new terms. The low realized productivity in spoken registers could mean there are less specific functions for this category in orally communicated language.

The second, hapax-based measure (expanding productivity) set a clear distinction between scientific writing and common language. The low expanding productivity in scientific writing could be explained by the need to repeatedly use created terms, instead of creating novel structures for single use. A high expanding productivity in spoken common language and fiction, in turn, means that the relative contribution of the suffix *-mine* to the growth rate of the whole vocabulary in these registers is higher and *-mine* nominalization is a likely category to attract novel concepts.

The third, category-internal measure (potential productivity) highlighted dialects as the most likely register for forming more types with the suffix *-mine* than are actually attested in the corpus, followed by fiction and spoken spontaneous speech. In other words, in CED, PCESS and the fiction subcorpus of BCE, the lexical variability in the category of *-mine* nouns appeared to be considerably higher than in the subcorpora of newspapers and scientific texts. This can be explained by dialectal vocabulary from 10 distinct dialect areas and creative language use in spoken language and fiction. As potential productivity is sensitive to opaque words, the lower potential productivity values in newspapers and scientific texts can also be a sign of a higher use of lexicalized and idiomatized *-mine* nouns in these registers. Examining the rate with which new types are formed with the suffix *-mine* in the corpora, thus, provides a clue for measuring the availability of this word-formation rule in different registers. The speed at which the lexical inventory of the types formed with *-mine* is enriched is considerably higher in spoken registers and thus, more linked with the syntactic, rather than lexical or textual functions of *-mine* derivation.

The role of lexicalized and idiomatized nouns in written registers was partly confirmed for newspaper texts based on NV-scores, which compared the ratios of verbal stems realized as verbs with the ratios of those stems realized as *-mine* nouns in the corpora. In newspapers, the stems most attracted to the nominalized construction were ones which are easily pluralized and used as referring to events or results of processes rather than to the processes themselves (*kohtumine* ‘meeting’, *valimine* ‘election’, *pakkumine* ‘offer’). However, the scores for these stems were high also in fiction and spoken spontaneous language, but not in scientific writing. In the latter, the stems which were far more likely to be nominalized

than used as verbs were the ones used in describing objects of study and scientific procedures (*käitumine* ‘behaviour’, *arutamine* ‘discussion’, *kasutamine* ‘using’, *kohaldamine* ‘customization’). In dialects, the verbal stems most attracted to *-mine* nominalization were words which described everyday activities (*söömine* ‘eating’, *elamine* ‘living’, *peksmine* ‘beating’). A common stem amongst all registers was *tege-* (‘do; make’) which occurs most often in specific grammatical constructions, namely the predicative construction in written language (*Tegemist on pettusega* ‘This is fraudery’) and the *busy*-construction in spoken language (*Emal oli tegemist* ‘Mother was busy’).

These results have to be accompanied by mentioning the possible pitfalls of this study. First, as many corpus studies, this study also suffers from known limitations, starting with the lack of control over the production to the non-representativeness of the subcorpora. The different registers in this study have been taken as predefined by the corpus compilation in order to provide sufficient data for quantitative analysis. However, it is clear that none of the five registers discussed are completely homogenous (or on the contrary, sufficiently heterogenous) in terms of the content and types of text they represent. Even scientific and academic writing have different, discipline- and publication-specific writing traditions, let alone media texts or spoken spontaneous language. Second, the measures used here do not account for the fact that the observations in the corpora are not independent, i. e. same writers/speakers can contribute multiple observations. Speakers of a linguistic community have individual preferences towards the grammatical means which the language provides and in order to reflect the preferences of the whole community, each speaker would ideally contribute only one observation or passage of equal length with other passages. This is often difficult to achieve with corpus studies due to the way in which the corpora have been compiled. Third, the methodology used in this article assesses the productivity of the suffix *-mine* only at the endpoint of sampling all the corpus files. This means that the non-linear growth curve of vocabulary is not accounted for. While this is somewhat less important in examining the productivity of one single suffix, this would have more elaborate consequences when comparing the productivity of different suffixes, since the productivity of less frequent suffixes could be overestimated.

Keeping all this in mind is necessary when drawing conclusions based on the results. However, the issues mentioned above do not lessen the importance of looking for usage-based evidence for morphological productivity of certain

derivational patterns. For quantitative studies of morphological productivity, corpus studies are shown to be more adequate than using dictionaries, and considerably faster than doing experimental studies, often providing a necessary starting point for the latter. The statistical productivity measures have several applications in e. g. authorship recognition, modeling processing constraints, and dictionary compilation.

There are plenty of alleys for further research. The continuation of the current study involves the comparison of the productivity measures for *-mine* with those of other deverbal action nominal suffixes, mainly *-us*, which is said to be less productive and semantically more complex. A clear issue with this is going to be the difficulty of obtaining appropriate data due to the morphological alternations. Based on the high degree of regularity, the productivity of *-mine* (*kõndimine* ‘walking’) could be compared with the productivity of the agent noun suffix *-ja* (*kõndija* ‘walker’). The variation in different registers should not be omitted from either study. The diachronic aspect, which has been neglected in this article, would shed light on the dynamics of change in the productivity of *-mine* in different registers, given that appropriate data can be found for also spoken language. Finally, the increasing amount of experimental research done in linguistics to complement corpus studies should not be disregarded in the study of morphological productivity of Estonian derivational affixes. Studies involving e. g. eyetracking experiments would significantly enrich our knowledge about the constraints involved in processing and producing examples of (non-)productive morphological categories.

7 Summary

This article described a study which assessed the morphological productivity of the Estonian deverbal suffix *-mine*, which is said to be the most regular and productive means for nominalizing a verb in Estonian. By examining corpus data representing five different registers of Estonian (written newspapers, fiction, scientific texts, spoken spontaneous common language, and spoken regional dialects) it was shown that the productivity of this suffix varies across the registers, depending on the specific aspects highlighted by 3 different productivity measures: realized productivity, expanding productivity, and potential productivity. Realized productivity, reflecting the contribution of *mine*-suffixed types to the size of the whole vocabulary, was highest in

written registers, especially in newspapers and scientific texts, while lowest in dialects. This is in accordance with the general claims about derivational patterns being much less productive in spoken registers than in written ones and is likely linked to the textual functions of *-mine* nominalization. Expanding productivity, which reflects the probability that it is indeed the suffix *-mine* that is used to grow the vocabulary by a type never encountered before, was highest in fiction and lowest in scientific texts, indicating that *-mine* is not a likely category for creating novel concepts in scientific texts, contrary to what intuition of a native speaker would imply. Finally, the third measure – potential productivity – characterizes *-mine* as being most productive in dialects and spoken common language. This means that spoken registers are lexically more diverse in the category of *-mine* nouns than would be expected by their overall lexical growth rate. In addition to these three measures, which showed that even a suffix with ‘absolute’ productivity exhibits variation in terms of its contribution to the vocabulary used in specific registers, the proportions of verb stems realized as verbs and as nominalized structures was compared in order to detect the stems with a higher likelihood of being more lexicalized. The stems attracted to the nominalized form were indeed the ones which have acquired additional meanings (e. g. *kohtumine* ‘meeting’, *valimised* ‘elections’) or are used in frequent grammaticalized constructions. While the samples used in this study were relatively small, the results provide a solid basis for more elaborate usage-based examination of derivational morphology in Estonian.

Acknowledgements

The study was supported by the European Union through the European Regional Development Fund (The Centre of Excellence in Estonian Studies).

Abbreviations

A	agent argument of transitive verb
ADV	adverbial
AN	agent noun
ADE	adessive case
GEN	genitive case
COM	comitative case
IPF	imperfect tense
IPS	impersonal mood
NM	nominalization (with the suffix <i>-mine</i>)
NOM	nominative case
PART	partitive case
POSTP	postposition
1INF	1st infinitive (the <i>-ma</i> infinitive)
2INF	2nd infinitive (the <i>-da</i> infinitive)
PPP	passive past participle
PL	plural
PRED	predicative
PRS	present tense
SG	singular
V	verb stem

References

- Anshen, Frank & Aronoff, Mark. 1999. Using dictionaries to study the mental lexicon. *Brain and Language* 68(1–2). 16–26.
- Baayen, R. Harald. 1992. Quantitative aspects of morphological productivity. In Booji, Geert E. & van Marle, Jaap (eds.), *Yearbook of morphology 1991*, 109–149. Dordrecht: Kluwer Academic Publishers.
- 1993. On frequency, transparency, and productivity. In Booji, Geert E. & van Marle, Jaap (eds.), *Yearbook of morphology 1992*, 181–208. Dordrecht: Kluwer Academic Publishers.
- 2003. Probabilistic approaches to morphology. In Bod, Rens & Hay, Jennifer & Jannedy, Stefanie (eds.), *Probabilistic linguistics*, 229–287. Cambridge, MA: The MIT Press.

- 2009. Corpus linguistics in morphology: Morphological productivity. In Lüdeling, Anke & Kytö, Merja (eds.), *Corpus linguistics: An international handbook*, 900–919. Berlin: Mouton de Gruyter.
- Baayen, R. Harald & Renouf, Antoinette. 1996. Chronicling The Times: Productive lexical innovations in an English newspaper. *Language* 72(1). 69–96.
- Baker, Paul & Hardie, Andrew & McEnery, Tony. 2006. *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.
- Barðdal, Jóhanna. 2008. *Productivity: Evidence from case and argument structure in Icelandic*. (Constructional approaches to language 8). Amsterdam: John Benjamins.
- Biber, Douglas. 1995. *Dimensions of register variation*. Cambridge: Cambridge University Press.
- Bybee, Joan L. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.
- Dressler, Wolfgang U. 2003. Degree of grammatical productivity in inflectional morphology. *Italian Journal of Linguistics* 15(1). 31–62.
- Erelt, Mati. 2017. Sekundaartarindiga laused [Sentences with secondary constructions]. In Erelt, Mati & Metslang, Helle (eds.), *Eesti keele süntaks* [The syntax of Estonian], 756–840. (Eesti keele Varamu III). Tartu: Tartu Ülikooli Kirjastus.
- Erelt, Mati & Kasik, Reet & Metslang, Helle & Rajandi, Henno & Ross, Kristiina & Saari, Henn & Tael, Kaja & Vare, Silvi. 1993. *Eesti keele grammatika II: Süntaks. Lisa: Kiri* [Estonian grammar II: Syntax. Appendix: Orthography]. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.
- Gaeta, Livio & Ricca, Davide. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* 44(1). 57–89.
- 2015. Productivity. In Müller, Peter O. & Ohnheiser, Ingeborg & Olsen, Susan & Rainer, Franz (eds.), *Word-formation: An international handbook of the languages of Europe*, vol. 2, 842–858. (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of linguistics and communication science (HSK) 40/2). Berlin: Walter de Gruyter.
- Hardie, Andrew & McEnery, Tony. 2006. Statistics. In Brown, Keith (ed.), *Encyclopaedia of language & linguistics*, 138–146. Oxford: Elsevier.
- Hay, Jennifer B. & Baayen, R. Harald. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences* 9(7). 342–348.
- Järvikivi, Juhani. 2003. *Allomorphy and morphological salience in the mental lexicon*. Joensuu: Joensuun yliopisto. (Doctoral dissertation).
- Kasik, Reet. 1968. Omastavalise täiendi subjektilisusest ja objektilisusest *mine*-konstruktsioonis [On the subjecthood and objecthood of the genitive attribute in *-mine* construction]. In *Keel ja struktuur: töid struktuuraalse ja*

- matemaatilise lingvistika alalt 2* [Language and structure: Works from structural and mathematical linguistics 2], 125–139. Tartu: Tartu Riiklik Ülikool.
- 1975. Verbi ja verbaalsubstantiivide tuletusvahekorrad tänapäeva eesti keeles [Derivational relations of verbs and verbal nouns in contemporary Estonian]. In *Keele modelleerimise probleeme 5* [Issues in language modeling 5], 4–162. (Tartu Riikliku Ülikooli toimetised 363). Tartu: Tartu Riiklik Ülikool.
- 1995. Õiguskeele lauseehitus [Syntax of legal texts]. *Õiguskeel* 1. 16–21.
- 2006. Nominalisatsioon meediauudiste tekstimoodustusvõttena [Nominalization as a textual strategy of news]. *Keel ja Kirjandus* 2/2006. 122–134.
- 2009. *Eesti keele sõnatuletus* [Derivation in Estonian]. Third, corrected edn. Tartu: Tartu Ülikooli Kirjastus.
- 2011. Sõnatuletus leksika ja grammatika vahel: *nd-* ja *ndus-*liitelised verbaalnoomenid [Derivation between lexicon and grammar: Verbal nouns with *-nd* and *-ndus* suffixes]. *Emakeele Seltsi aastaraamat* 56 (2010). 63–90.
- 2013. *Komplekssete sõnade struktuur* [Structure of complex words]. (Tartu Ülikooli eesti keele osakonna preprintid 3). Tartu: Tartu Ülikool.
- Kerge, Krista. 2001. Action nouns on *-mine* in Estonian grammar and lexicon. In Niemi, Jarkko & Heikkinen, Janne (eds.), *Nordic and Baltic morphology: Papers from a NorFa course, Tartu, June 2000*, 34–45. (Studies in Languages 36). Joensuu: University of Joensuu.
- 2002. Kirjakeele kasutusvaldkondade süntaktiline keerukus [Syntactic complexity of functional domains of literary language]. In Kasik, Reet (ed.), *Tekstid ja taustad: Artikleid tekstianalüüsist* [Texts and backgrounds: Articles on text analysis], 29–46. (Tartu Ülikooli eesti keele õppetooli toimetised). Tartu: Tartu Ülikooli Kirjastus.
- 2003. *Keele variatiivsus ja mine-tuletus allkeelte süntaktilise keerukuse tegurina* [Linguistic variability and *-mine* derivation as a factor in lectal syntactic complexity]. Tallinn: TPÜ kirjastus. (Doctoral dissertation).
- Mäkisalo, Jukka. 2000. *Grammar and experimental evidence in Finnish compounds*. Joensuu: Joensuun yliopisto. (Doctoral dissertation).
- Neetar, Helmi. 1988. Mõnest teonimekonstruktsioonist Eesti murretes [On some action nominal constructions in Estonian dialects]. *Emakeele Seltsi aastaraamat* 32 (1986). 36–45.
- Nikolaev, Alexandre & Niemi, Jussi. 2008. Suomen nominien taivutusjärjestelmän produktiivisuuden indekseistä [On the productivity indices of the inflectional paradigm of Finnish nouns]. *Virittäjä* 112(4). 518–544.
- Pilvik, Maarja-Liisa. 2016. *olema + Vmine* konstruktsioonid eesti murretes [Constructions with *olema + Vmine* in Estonian dialects]. *Keel ja Kirjandus* 6/2016. 429–446.

- 2017. Deverbal *-mine* action nominals in the Estonian Dialect Corpus. *Eesti ja soome-ugri keeleteaduse ajakiri ESUKA / Journal of Estonian and Finno-Ugric Linguistics JEFUL* 8(2). 295–326.
- Pinker, Steven. 1999. *Words and rules: The ingredients of language*. New York: Basic Books.
- Plag, Ingo. 1999. *Morphological productivity: Structural constraints in English derivation*. Berlin: Mouton de Gruyter.
- Plag, Ingo & Dalton-Puffer, Christiane & Baayen, R. Harald. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3(2). 209–228.
- Rainer, Franz. 1987. Produktivitätsbegriffe in der Wortbildungslehre [Productivity notions in word-formation]. In Wolf, Dietrich & Gauger, Hans-Martin & Geckeler, Horst (eds.), *Grammatik und Wortbildung romanischer Sprachen* [Grammar and word formation in the Romance languages], 187–202. Tübingen: Narr.
- Sahkai, Heete. 2011. *Teine grammatika: Eesti keele teonimede süntaks konstruktsioonipõhises perspektiivis* [Second grammar: The syntax of Estonian action nominals in a constructionalist perspective]. Tallinn: Tallinna Ülikool. (Doctoral dissertation).
- Vare, Silvi. 1994. *Nimi- ja omadussõnatuletus tänapäeva kirjakeeles* [Nominal and adjectival derivation in contemporary literary language]. Tartu: Tartu Ülikool. (Doctoral dissertation).

Contact information:

Maarja-Liisa Pilvik
 Institute of Estonian and General Linguistics
 University of Tartu
 Jakobi 2
 51014 Tartu
 Estonia
 e-mail: maarja-liisa(dot)pilvik(at)ut(dot)ee

The pronoun *kes* ‘who’ and its referent’s animacy in Estonian dialects

Hanna Pook
University of Tartu

Abstract

In the Estonian language the interrogative-relative pronouns *kes* ‘who’ and *mis* ‘what’ are distinguished by whether they are used to refer to animate or inanimate entities, respectively. However, in certain Estonian dialects the pronoun *kes* can be used to refer to inanimate entities as well. The aim of this paper is to quantitatively analyse this non-standard use of *kes* based on the data in the Corpus of Estonian Dialects and to determine which dialects typically use *kes* to refer to inanimate entities and which variables (construction type, case marking, number, relative clause position and referential distance) contribute most to this phenomenon. The results show that there are significant differences between the dialects – the pronoun *kes* is mostly used to refer to inanimate entities in Northern Estonia (most frequently in the Eastern dialect, but also in the Mid, Coastal and Western dialects), but this phenomenon is rare in the southern dialects. This paper argues that out of all the variables, the choice to use *kes* to refer to inanimates is most associated with the case of the pronoun, being very frequent in elative and comitative. It also appears that *kes* is used to refer to inanimate entities mostly in postnominal relative clauses and when the word expressing the referent is in plural.

Keywords: dialect syntax, Estonian dialects, interrogative-relative pronouns, animacy

1 Introduction

In many languages there seems to be a strong connection between the animacy of a referent and the choice of ways to refer to it (Dahl & Fraurud 1996: 56). This is also true for the interrogative-relative pronouns *kes* ‘who’ and *mis* ‘what’ in Standard Estonian: *kes* is usually used to refer to animate entities and the pronoun *mis* is used to refer to inanimate entities.

However, in some Estonian dialects the pronoun *kes* can be used to refer to inanimate entities (examples 1–2)¹ and *mis* can be used to refer to animate entities.

- (1) Eastern dialect (Avinurme)²

sell ael eij õld jällä noh seda saage kellega
 that:ADE time:ADE not be:PST:PTCL again PART this:PRT saw:PRT who:COM
puid leigatta
 tree:PL:PRT cut:INF

‘At that time there was no saw to cut trees with.’

- (2) Insular dialect (Käina)

narid old kaa kelle sees magadi
 bunk:PL be:PST:PTCL too who:GEN in sleep:IPS:PST

‘[There] were bunkbeds, too, that were slept in.’

There has been little written about this possibility in previous research. Kask & Palmeos (1985: 59) mention that in some areas of Estonia the relative pronoun *kes* can be used to talk about inanimate entities. This phenomenon is also noted in a few descriptions of certain Estonian subdialects (see Niklus 1957: 139; Pung 1968: 134; Juhkam 2012: 468). However, these previous works only mention the existence of this phenomenon, but do not elaborate on how or why *kes* is used in this way.

The aim of this article is to give a thorough overview of this non-standard use of *kes* for referring to inanimate entities in the Estonian dialects based on the Corpus of Estonian Dialects.³ The research questions are the following:

1. Which Estonian dialects and subdialects typically refer to inanimate entities with *kes*?
2. Do construction type, the case of the pronoun and the number of the referent affect whether inanimate entities are referred to with *kes*, and if so, how?

¹ These and all the following examples are from the Corpus of Estonian Dialects. A brief description of the corpus is given in § 2.1.

² Each example is preceded by the dialect and subdialect (in parentheses). The description and map of Estonian dialects and subdialects is given in § 2.1.

³ This article is largely based on the author’s master’s thesis *Pronoomeni kes kasutusest eesti murretes* (‘The use of the pronoun *kes* ‘who’ in Estonian dialects’, Pook 2018), with little corrections and an extended analysis.

3. What is the difference in the use of the pronoun *kes* between pre- and postnominal relative clauses?
4. Is the referential distance between the pronoun *kes* and the referred to entity in the relative clause connected to the animacy of the referent?

This paper is structured as follows: in § 1.1 I give an overview of how animacy is conveyed in languages and in § 1.2 I explain the principles guiding the choice between *kes* and *mis* in Estonian. In § 2 I describe the Estonian dialects, the data found in this study and the methods used. The following analysis of the data is separated into two parts. § 3 describes the association between the analysed variables and the animacy of the referent to which the pronoun *kes* refers, presents the results of this analysis and answers research questions 1 and 2. In § 4 I make a subset of the data containing only those dialect areas that typically use *kes* to refer to inanimate entities (based on the results of § 3) and only headed relative clauses, and analyse those separately. Therefore, § 4 answers research questions 3 and 4. § 5 contains the conclusions and discussion.

1.1 Animacy

Animacy in a language is hardly ever a binary variable like in biology, but regarded more as a continuum that extends from human through animal to inanimate (Yamamoto 1999: 1). The first modern description of this continuum, called the (universal) animacy hierarchy, was presented by Silverstein (1976), but the most common representation of this hierarchy, found in Dixon (1979: 85), is as follows:

1st, 2nd person pronoun > 3rd person pronoun > proper names >
human common noun > non-human animate common noun > inanimate
common noun

Since then, this hierarchy has been used and adapted by countless linguists to describe different construction types in different languages. Some languages and/or constructions use less fine distinctions, e.g. just human versus non-human or animate versus inanimate, while others can have intermediate categories between the levels of the common hierarchy, e.g. humans > higher animals > nonhuman animates or have 1st and 2nd person as separate categories (Dixon 1979: 85; Comrie 1989: 185; Croft 1990: 113).

The categories that are higher in the hierarchy are often grammatically distinguished from those that are lower, are treated as more central to clause structure and are more likely to act as an agent in events (Whaley 1996: 172; Kittilä et al. 2011: 6). Still, this hierarchy is definitely not an ordering of discrete categories, but rather a continuum ranging from most to least animate (Croft 1990: 113).

However, animacy is not the only important parameter reflected in this hierarchy – in fact, there are several others. One of them is empathy: hierarchy assigns primacy to nominals with which the speaker is familiar or has empathy towards (Whaley 1996: 172–173). Certain authors prefer to avoid the term *animacy* altogether and speak rather of *empathy hierarchies*, since there is technically no difference in literal animacy between the 1st person and a proper noun or a human common noun (Kittilä et al. 2011: 6; see also Langacker 1991: 306–307). The empathy hierarchy according to Langacker is the following:

speaker > hearer > human > animal > physical object > abstract entity

Another relevant parameter to the animacy hierarchy is the individuation scale: a stronger sense of animacy is attributed to entities that are seen as individuals than to those that are part of an indeterminate mass (Comrie 1989: 189). The most important distinction in individuality is between singularity and plurality, but there are several examples of ‘masses’ that lie on the boundary between animacy and inanimacy, e. g. human organisations, geographic entities or communities, which are usually treated as inanimate, but still show certain animate-like features (Yamamoto 1999: 131, 138–139).

All in all, determining whether something is animate or inanimate is not always clear. Not even all animate beings are equally animate to human cognition, since animacy has a gradience from centrally animate beings to peripherally animate beings (Yamamoto 1999: 14). According to Fowler (1977: 16–17), animate beings are those that are capable of initiating actions and change, either through conscious or unconscious drive (whereas inanimate beings lack this capability to cause the world to change), and of movement (which is why plants are typically perceived as inanimate). In this study I follow Fowler’s distinction, categorising all humans (and human collectives) as well as animals as animate. Therefore this paper mainly focuses on the animacy hierarchy instead of the empathy hierarchy, since it is generally possible to categorise the entities in the dataset by their biological animacy. A more detailed description of the data coding can be found in § 2.2.

1.2 *kes* and *mis* in Estonian

According to Erelt (2017: 743), the choice between *kes* ‘who’ and *mis* ‘what’ in Standard Estonian is determined by the animacy hierarchy:

humans > higher animals > lower animals⁴ > inanimates

and the individuation scale:

an individual > a distributive group > a collective

With these hierarchies, the probability of using *kes* decreases and the probability of using *mis* increases from left to right. The relative pronoun *kes* is always (in Standard Estonian) used for a single human referent and the relative pronoun *mis* is always used to refer to an inanimate entity. In the case of multiple referents, *kes* is preferred for a distributive group and *mis* is preferred for a collective group. The choice of the pronoun also depends on whether the collective or institution is in the role of agent or patient – in the first case, *kes* is preferred, while in the second case, *mis* is preferred. It can be generalised that the more active the role of the referred to entity, the more likely the use of the pronoun *kes* is (Erelt 2014: 743).

Animate non-human entities can be referred to using both *kes* and *mis*. The pronoun *mis* can even be used to refer to a single animal: with higher animals, the use of *mis* is justified if the animal is not the agent in the sentence, for lower animals the choice of the pronoun is freer and it is always possible to use the pronoun *mis* (Erelt et al. 2007: 561).

Examples that run contrary to these principles have not only been found in Estonian dialects, but are also apparent in old literary Estonian: while in those texts *kes* usually refers to animate entities, there are certain relative clauses where the relative pronoun *kes* refers to inanimate referents, such as *usk* ‘belief’, *puu* ‘tree’, *kiri* ‘letter’, etc. (Kõpp 2001: 41–42). According to Erelt (1996: 11) the choice between the pronouns *kes* and *mis* has not always functioned according to contemporary norms, even in the written texts of the 19th and beginning of the 20th century.

In addition, Wiedemann (2011: 470) notes in his Estonian grammar written in 1875 that in Mid-Estonia only the interrogative pronoun *kes* (referring to animates) was carefully kept apart from the interrogative pronoun *mis* (referring to inanimates), while using these pronouns in a relative clause, a

⁴ Erelt does not specify what he considers to be a higher or a lower animal.

similar distinction was often not made (however, he also remarks that in some dialects the distinction was not made for interrogative pronouns either).

It is possible that a clear distinction of interrogative pronouns, which was not followed as precisely with relative pronouns, is related to the grammaticalisation of the interrogative marker. In Estonian (and in most Finno-Ugric languages) interrogative sentences are an older construction than relative clauses, which means that the interrogative pronouns are older than the relative pronouns (Alvre 1987: 23). According to Heine & Kuteva (2006: 209), interrogative markers have four possible stages of grammaticalisation, from being just an interrogative marker to introducing headed relative clauses:

1. Only an interrogative marker (*Who came?*).
2. Interrogative marker is extended to introduce indefinite complements or adverbial clauses (*I don't know who came.*).
3. In addition to indefinite clauses, the interrogative marker can introduce definite complement or adverbial clauses, which can also be interpreted as headless relative clauses (*You also know who came.*).
4. In addition to definite non-headed relative clauses, the interrogative marker can introduce headed relative clauses (*Do you know the woman who came?*).

Because the interrogative marker in Estonian (and in most Balto-Finnic languages) has gone through all of these grammaticalisation stages, the words for both interrogative and relative pronouns are the same (Erelt 2017: 683). It has been noted about many languages that going down the stages decreases the distinctions in gender, animacy, number and case and the marker assumes a more general (grammatical) meaning (Heine & Kuteva 2006: 227).

Based on that, we can deduce that in Estonian the pronouns *kes* 'who' and *mis* 'what' are the most systematically distinguished by animacy in independent interrogative sentences and the least in relative clauses. Therefore, relative clauses could be the construction type where *kes* most probably refers to inanimate entities, because in such constructions the meaning of the interrogative marker is probably the most abstract and unclear. In this paper I attempt to find evidence to support this theory.⁵

⁵ It should be noted that *kes* and *mis* are not the only pronouns that have lost some of their

2 Data and methods

2.1 Estonian dialects, Corpus of Estonian Dialects and data

The analysed data has been collected from the Corpus of Estonian Dialects (CED),⁶ which contains authentic spoken texts from all Estonian dialects. There are several different ways that the dialects can be divided and grouped, but in this corpus they have been divided into ten traditional dialect areas: the Mid, Western, Insular and Eastern dialects (i. e. the North-Estonian dialect group), the Coastal and Northeastern dialects (i. e. the Northeastern-Coastal dialect group), and the Tartu, Võru, Mulgi and Seto dialects (i. e. the South-Estonian dialect group) (Lindström 2015). These dialect areas can be seen in Figure 1.

The most significant contrast exists between the northern (North-Estonian and Northeastern-Coastal dialect groups) and southern dialects, with the greatest differences being found in their phonology, morphology and lexis. According to K. Pajusalu et al. (2009: 73), the oldest distinctive features of these two dialect regions can be traced back thousands of years, when they were the differences between Balto-Finnic tribal languages. Still, most of the distinct features developed later due to language contacts, administrative borders, etc. South-Estonian dialects are at times even considered to be a separate language. Recent studies in dialect syntax have, however, found that for certain constructions the main isogloss line can be drawn rather between the western and eastern dialects (see e. g. Uiboed 2013; Uiboed et al. 2013; Lindström et al. 2014; 2015; 2018).

All the dialects in the corpus have been divided into subdialects (the borders of which are based on former administrative units – parishes). The subdialects represented in the corpus data that is used in the present article are shown in Figure 2. It is clear that the representation of subdialects differs between dialects – e. g. the data contains all of the subdialects of the Northeastern and Võru dialect, but only about half of the subdialects are

distinction in animacy in spoken Estonian. In Standard Estonian, the personal pronoun *tema/ta* ‘s/he’ is used to refer to animate entities and the demonstrative pronoun *see* ‘it’ is used to refer to inanimates. However, in spoken Estonian, in the case of a very familiar entity or in some idiomatic expressions, *ta* (the short form of the pronoun) is often used to refer to inanimates. Similarly, the demonstrative *see* is used to refer to animate entities either in a presentational clause, when there are two different animate referents involved or for emphasis in the case of a relatively new referent (R. Pajusalu 2005: 110, 132–133).

⁶ <http://www.murre.ut.ee/estonian-dialect-corpus/> (accessed 2020-04-16).

Table 1. The number of the informants, total tokens and keyword *kes* in the dialects

dialect	informants	total tokens	keyword <i>kes</i>
Western	62	251,031	741
Mid	70	246,167	680
Insular	53	202,325	621
Võru	33	111,503	389
Coastal	24	97,152	329
Eastern	20	48,353	205
Seto	23	68,414	195
Tartu	21	80,343	191
Mulgi	16	63,759	187
Northeastern	15	60,037	171
Σ	337	1,229,084	3,709

represented in the Mid dialect. It should be noted, however, that the Mid dialect area is one of the biggest dialect areas, so there are still more texts from that dialect (and other bigger dialects that are not represented in this data in their entirety) than from, for example, the Northeastern or Võru dialect.

The CED includes dialect recordings, phonetically transcribed texts, dialect texts in simplified transcriptions, morphologically annotated texts and a separate database for information about the informants and recordings. This analysis uses (manually) morphologically annotated dialect texts, from which all the uses of the pronoun *kes* (based on the lemma ‘kes’) have been collected into the data, in total 3,709 keywords. Every keyword is accompanied by the preceding and following context (up to 20 words), case marking (already annotated in the corpus texts) and information about the informant (dialect, subdialect, village, age, date of birth, gender and place of birth).

The frequency of the keyword *kes* in different dialects is presented in Table 1. The table also includes the number of informants and the total number of tokens in the texts from each dialect. As expected, the number of informants and the length of the texts mostly correspond with the frequency of the keyword. The number of the informants and total tokens varies so much across the dialects mainly because some dialects areas are much bigger than others and some dialects have had more available informants.

Out of the 337 informants 236 were women and 101 were men. They were

born between 1864 and 1922 and ranged in age from 49 to 100 years old. The texts were recorded mostly in the 1960s and 1970s, but the earliest recordings date to the year 1938 (Lindström 2015). This means that the language analysed in this paper is antiquated (although the dialects are more or less still in use in the rural areas of Estonia, especially in Southern Estonia and in the insular region).

2.2 Data coding

The data attained from the corpus contains, as mentioned above, keywords, context, case and information about the informant. In addition, I have manually added certain variables. I did this in two stages. In the first stage I added the referent's animacy, construction type and number of the word expressing the referent to every sentence in the data. I also labelled whether relative clauses were headed or non-headed.

Animacy's variable has two levels: animate and inanimate. As previously mentioned, I have followed Fowler's (1977: 16–17) distinction in determining animacy: all humans (and human collectives) as well as animals are categorised as animate, while everything else is categorised as inanimate. Due to the nature of the analysed texts in the CED (which typically include topics like the informant's personal life, their lifestyle, past events or working methods), the marking of the referred to entities as animate or inanimate was relatively straightforward and there were hardly any borderline cases of animacy.

The sentences were divided into seven construction types based on which types most frequently appeared in the data. A detailed description of all the construction types is presented in § 3.2.

The number of the word expressing the referent is either singular or plural. (It should be noted that this is not the number of the pronoun, since the pronoun *kes* does not have plural forms and is only used in singular.)

In the second stage, in the second part of the analysis, I added two variables only to headed relative clauses: distinguishing whether the relative clause is pre- or postnominal as well as measuring the distance in words between the pronoun *kes* and the word expressing the referent. The specifics of those variables are explained thoroughly in § 4.1 and § 4.2. All the variables, their levels and abbreviations are presented in Table 2.

The CED consists of spoken texts, which means a significant number of sentences in the dataset are incomplete or contain repetition of the pronoun

Table 2. The variables in the dataset and their levels. If applicable, the abbreviations of the levels used in subsequent graphs are given in parentheses.

variable	levels
dialect	Eastern (E), Western (W), Mid (MID), Insular (I), Coastal (C), Northeastern (NE), Mulgi (MUL), Tartu (T), Võru (V), Seto (S)
animacy	animate (anim), inanimate (inanim)
construction type (TYPE)	relative clause (rel), relative clause without a main clause (rel_main), question (q), indirect question (indir_q), rhetorical question (rhet_q), listing construction (list), other
case marking (CASE)	nominative (nom), genitive (gen), partitive (prt), elative (ela), allative (all), adessive (ade), translative (trl), comitative (com)
number (NUM)	singular (S), plural (P)
relative clause position (REL_POS)	prenominal (pre), postnominal (post)
referential distance in words (DISTANCE)	0, 1, 2, 3, 4, 5, 6, 7+

kes. Since it was not possible to analyse unfinished sentences and including repetitions did not add anything to the analysis, these sentences were excluded from the dataset.

In the end, the final dataset consisted of 3,324 keywords *kes* with context.

2.3 Methods

This study is a corpus-based dialect analysis, which has many advantages over classic atlas-based or dictionary-based approaches, some of which I will expand on here. The corpus is based on natural (spoken) language use, so the data from the corpus is authentic and reliable (whereas atlas data can sometimes be biased, since it has mostly been collected by questionnaires and therefore can only represent conscious language use). Corpus-based dialectology takes into account the variation and the frequency of different variants, so instead of being able to say that a certain feature exists in one dialect and does not exist in another, we can calculate the exact frequencies

and differences between dialects. The corpora also include the immediate context around the feature, which gives an opportunity to analyse semantic, pragmatic and textual characteristics of the feature (Lindström & Pilvik 2018: 649).

There are also some restrictions to the data when using corpus-based approaches (see Szmrecsanyi & Anderwald 2018: 302), but overall the possibility to systematically look at the features and the extent of their usage gives us a much wider picture of the variation than previous dialect atlas data could.

To determine which variables most influence the decision to use the pronoun *kes* to refer to inanimate referents, I have analysed the data in two ways: firstly I have looked at the results quantitatively, inspecting the frequencies of the values and (if possible) finding significant correlations between them using the Chi-Squared test. Secondly, I have applied the conditional inference tree method. This was chosen due to its suitability for analysing the data, which contains mostly nominal variables that are often not acceptable as inputs using other statistical methods. Unlike parametric tests, this method does not require any distributional assumptions to be met (which this dataset does not). Conditional inference trees (and random forests) have also been successfully used in other studies about Estonian dialect syntax (see Klavan et al. 2015; Ruutma et al. 2016; Lindström & Uiboed 2017; Lindström et al. 2018, etc.).

The conditional inference tree method is based on binary recursive partitioning, wherein each stage the algorithm tests whether any of the independent variables are associated with the given response variable. The variable that is most strongly associated with the response is selected for the next split, which divides the dataset into two subsets. This continues until there are no variables that are associated with the response at the level of statistical significance. The result of this process is depicted as a tree structure with binary splits (Levshina 2015: 291).

All the calculations for the results were performed using the program R (R Core Team 2018). The conditional inference trees were computed using the function `ctree()` in the package `party` (Hothorn et al. 2006).

3 Impact of the analysed variables

In this section of the paper, I analyse different variables to find out how they affect the decision to use the pronoun *kes* to refer to inanimate entities. The analysis has been divided into two parts. In this section (§ 3) I examine how the animacy of the referent is associated with dialects and subdialects, different construction types, the case of the pronoun and the number of the referent.

In the following section (§ 4) I extract a subset from the data that only consists of headed relative clauses and those dialect areas that (according to the results of § 3) use *kes* frequently to refer to inanimate entities, to see whether pre- and postnominal relative clauses or the referential distance between the pronoun and the word used to express the referred to entity have any differences in regards to the usage of the pronoun *kes*.

3.1 Dialects and subdialects

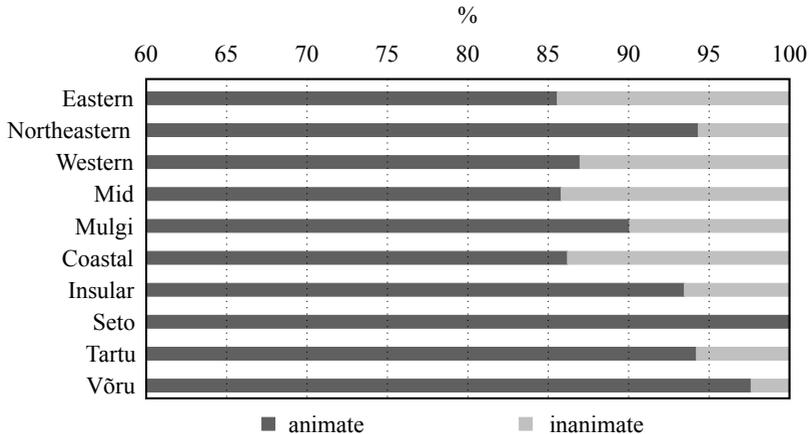
In order to find the differences in the use of the pronoun *kes* between the dialects, I compared the dialects and subdialects represented in the data in terms of how frequently inanimate entities were referred to using the pronoun *kes*. The hypothesis was that referral to inanimate entities using *kes* is typical in the North-Estonian dialect group, particularly in the Mid and Western dialects, but does not occur in the South-Estonian dialect group. This hypothesis was based on a previous automated study of the same data, which provided adequate results to establish this hypothesis. However, that study made it clear that an automated analysis of spoken text cannot give conclusive results. Therefore, for this study all the material was analysed manually.

The frequency of animate and inanimate referents that the pronoun *kes* is referring to is separated by dialect in Table 3. The percentages of animate and inanimate referents are presented in Figure 3. Using the pronoun *kes* to refer to inanimate entities is most frequent in the Eastern (14.4% of all the pronoun *kes* occurrences were used to refer to an inanimate entity, 28 pronouns out of 194), Mid (14.2%, 87 out of 613), Coastal (13.8%, 41 out of 297) and Western dialects (13%, 88 out of 676). The feature is quite rare in the Tartu (5.8%, 10 out of 173), Northeastern (5.7%, 9 out of 159) and Võru dialects (2.4%, 8 out of 336), and it does not occur at all in the Seto dialect.

This confirms the hypothesis that referring to inanimate entities with *kes* is more frequent in the North-Estonian dialect group, but the percentages are clearly not homogenous within all the dialect groups. The Insular

Table 3. The raw frequency of animate and inanimate referents in dialects

dialect	animate	inanimate	Σ
Mid	526	87	613
Western	588	88	676
Insular	515	36	551
Eastern	166	28	194
Coastal	256	41	297
Northeastern	150	9	159
Mulgi	145	16	161
Tartu	163	10	173
Võru	328	8	336
Seto	164	0	164
Σ	3,001	323	3,324

**Figure 3.** The percentage of animate and inanimate entities referred to by the pronoun *kes* in different Estonian dialects

dialect differs the most from the rest of the North-Eastern dialect group. Moreover, the Coastal and Northeastern dialects, which both belong to the Northeastern-Coastal dialect group (and are even categorised as one Coastal dialect in the Estonian Dialect Dictionary⁷), differ significantly in regard to this particular phenomenon. Therefore, it might not be possible to completely categorise this use of *kes* according to the traditional dialect groups.

Nevertheless, we can say that the South-Estonian dialect group stands out the most from the other groups. It can thus be concluded that in the South-Estonian dialects the pronoun *kes* is used differently from other dialect groups and the referred to entity is almost always animate. The examples in the data where *kes* refers to inanimate entities in those dialects are more exceptions than a typical use of the pronoun. There were significantly more inanimate referents in the Mulgi dialect than in other South-Estonian dialects, but those were mostly due to the idiolect of one informant. However, other studies in dialect syntax have found that the Mulgi dialect cannot always be grouped together with the other southern dialects, but rather behaves more similarly to the northern dialects (see Lindström et al. 2015; Ruutma 2016), which might also be the reason for the Mulgi dialect’s distinctiveness.

In the following map of all the subdialects and the percentage of the referenced inanimate entities (Figure 4), it can be seen that the use of the pronoun *kes* to refer to inanimate entities is most frequent in Northern and Central Estonia, but is considerably less frequent in Southern and Eastern Estonia and on the island of Saaremaa. However, based on the subdialects represented in the data, there does not seem to be a specific dialect area even in Northern Estonia where this use of *kes* is more prevalent than elsewhere. Instead, there are single subdialects (or groups of subdialects) all over Northern Estonia that tend to use this feature more, bordered by subdialects that rarely or never use *kes* to refer to inanimate entities.

Looking at the subdialects more closely, it appears that using *kes* to refer to inanimate entities is mostly widespread all over the Eastern, Coastal, Western and Mid dialects, with very few exceptions. The subdialects in which the most inanimate entities were referred to using *kes* were Avinurme and Palamuse from the Eastern dialect (20% [6 pronouns out of 30] and 19.1% [16 out of 84] of all the referenced entities were inanimate, respectively), Juuru and Koeru from the Mid dialect (46.9% [15 out of 32] and 24% [12 out of 50], respectively), Haljala and Jõelähtme from the Coastal dialect (26.2% [11 out

⁷ <http://www.eki.ee/dict/ems/> (accessed 2020-04-16).

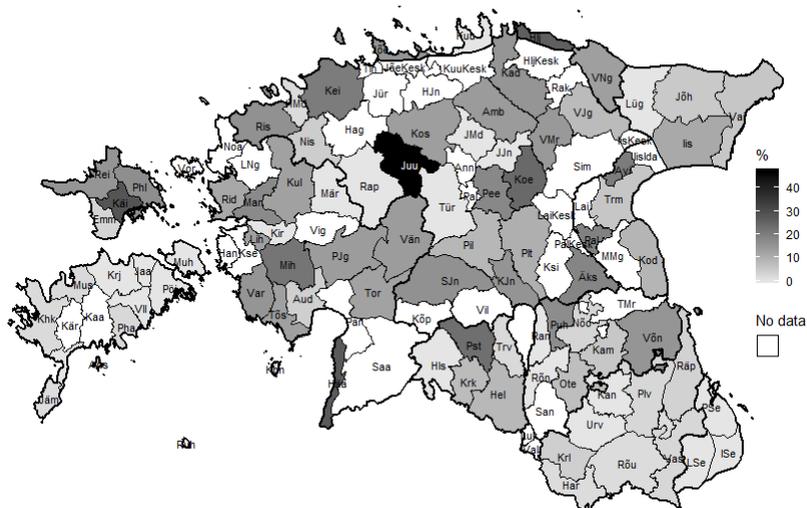


Figure 4. The percentage of the pronoun *kes* used to refer to inanimate entities in the represented subdialects. The thicker lines represent the borders between dialects.

of 42] and 15.1% [18 out of 119], respectively) and Häädemeeste and Mihkli from the Western dialect (27.3% [6 out of 22] and 21.4% [6 out of 28], respectively).

Even though the average percentage of inanimate referents is low in the Insular dialect (only 6.5%, 36 pronouns out of 551) and most subdialects do not refer to an inanimate entity with *kes* at all, there are three subdialects that considerably raise the overall percentage of the dialect: Käina (28.6%, 18 out of 63), Pühalepa (16.4%, 10 out of 61) and Reigi (15%, 3 out of 20). All these subdialects are spoken on the island of Hiiumaa. Therefore, while the rest of the Insular dialect uses *kes* more like the South-Estonian dialect group, Hiiumaa's subdialects clearly have the option of using *kes* to refer to inanimate entities.

The reason for this is not clear – it is possible that using *kes* to refer both to animate and inanimate entities in those subdialects has been influenced by contacts with the Swedish language, which uses one main relative pronoun

som to refer to both animate and inanimate entities (Tuldava 1993: 144–147). Inanimate entities in Hiiumaa’s subdialects are mainly referred to by *kes* in relative clauses, so this could be one explanation for the frequent use of *kes*.

However, language contact is a wide topic and its effect is difficult to measure without a larger study. In a previous study of mine (Pook 2018) one of the research questions was whether some of the contact or cognate languages have influenced the use of the pronoun *kes* in Estonian dialects, but it was not possible to come to any clear conclusions about that. Therefore, the possible effect of language contact is excluded from the scope of this paper.

It should be noted, nevertheless, that language contact has been said to affect and even accelerate grammaticalisation (Heine & Kuteva 2005: 14), which, as described in § 1.2, is one likely reason behind the phenomena studied in this paper. For example, language contact induced grammaticalisation has played an important role in the development of articles in Eastern European languages, including the article-like demonstrative pronoun *see* ‘it’ and numeral *üks* ‘one’ in Estonian language (Hint et al. 2017: 68–69; see also R. Pajusalu 1997; 2009; Heine & Kuteva 2006). So while language contact is not discussed in this paper, upon further reading one must keep in mind that it may not only be a possible reason behind the pronoun *kes* being used to refer to inanimates, but it may also be a hidden factor behind the grammaticalisation of the pronoun.

3.2 Construction type

In this section, I look at how the factor of construction type is related to the informants’ choice to use the pronoun *kes* for referring to inanimate entities. I divided all the sentences into seven types, basing the division on types most frequently represented in the data. These included very common construction types like relative clauses (3), interrogative sentences (5), rhetorical questions (6) and indirect questions (7), but also some that are unique to the data or to the spoken language, such as relative clauses lacking a main clause that it is supposed to be modifying (4; these kinds of constructions either had the function of introducing a new topic or lacked a main clause simply because the speaker’s train of thought changed mid-sentence), repeated referencing while listing (8; repeatedly using *kes* to refer to a part of a group) and other constructions that did not belong to any of the previous types.⁸

⁸ The entities referred to by *kes* in these examples are all animate, since the purpose of them is to illustrate construction types and not the studied phenomena.

- (3) Northeastern dialect (Jõhvi)

igal maeal kaks ärga kes olid künni jaust
 every:ADE house:ADE two ox:PRT who be:PST:3PL plow:GEN for
 ‘Every house [had] two oxen that were for plowing.’

- (4) Mid dialect (Põltsamaa)

kell ol maa nigu mull ol maa mull
 who:ADE have:PST:3SG land like me:ADE have:PST:3SG land me:ADE
ess ole tuuga aigugi tettä ess
 not:PST have:CNG that:COM time:PRT:CLI do:INF not:PST
 ‘[S/he] who had land like I had land, I did not even have time to do [anything] with that.’

- (5) Western dialect (Martna)

kelle müts oli siis see selle Lauri vana müts või
 who:GEN hat be:PST:3SG then this that:GEN Lauri:GEN old hat or
 ‘Whose hat was this, that Lauri’s old hat?’

- (6) Insular dialect (Muhu)

nee oo meelest juba kadun kiss teab kus
 that:PL be:3SG mind:ELA already disappear:PST:PTCL who know:3SG where
naad läin oo
 they go:PST:PTCL be:3SG
 ‘Those are gone from memory already, who knows where they have gone.’

- (7) Võru dialect (Põlva)

eit tiijäq kess tu pesä sõss ol maaha ajanu
 not know:CNG who that nest:GEN then be:PST:3SG down push:PST:PTCL
 ‘[I] don’t know who had pushed that nest onto the ground.’

- (8) Seto dialect (Eastern-Seto)

kess kudi kinnast kes kudi sukka
 who knit:PST:3SG glove:PRT who knit:PST:3SG stocking:PRT
 ‘Who knitted a glove, who knitted stockings.’

The hypothesis was that referring to inanimates with *kes* is most common in relative clauses, since in the process of the interrogative markers’

grammaticalisation (from being just an interrogative marker to being able to introduce headed relative clauses) they could have lost their distinction in animacy, as described by Heine & Kuteva (2006: 209, 227). The detailed reasoning behind this has already been explained in § 1.2.

We can see in Table 4 that the majority of the data consists of relative clauses: 71.3% of the constructions are relative clauses.

Relative clauses also include the highest percentage of inanimate entities referred to by *kes* (12.4%), confirming my hypothesis. Pronouns in other construction types were rarely (or not at all) used to refer to inanimate entities. The percentage of the pronoun *kes* being used to refer to inanimate entities by construction type is depicted in Figure 5.

The data contained a significant number of non-headed relative clauses (i. e. relative clauses lacking a head noun they were modifying; example 3 headed, example 9 non-headed). Despite that omitting the head is grammatically incorrect in Standard Estonian, it is quite frequent in spoken language. There were 341 non-headed relative clauses in the data, of which only 10 contained an inanimate entity (10–11).

(9) Mulgi dialect (Helme)

kes tegi parembat rohkemp tüüd sai ka rohkemp
 who do:PST:3SG good:CMP:PRT more work:PRT get:PST:3SG also more
palkka
 pay:PRT

‘[S/he] who did better [and] more work also got more pay.’

(10) Mid dialect (Suure-Jaani)

omale jäi kaa keda sa jäll tuleva
 self:ALL remain:PST:3SG too who:PRT you again come:PTCL:GEN
kevadi maha külisid
 spring:GEN down sow:PST:2SG

‘[Some] was left for you as well that you sowed again next spring.’

(11) Western dialect (Tõstamaa)

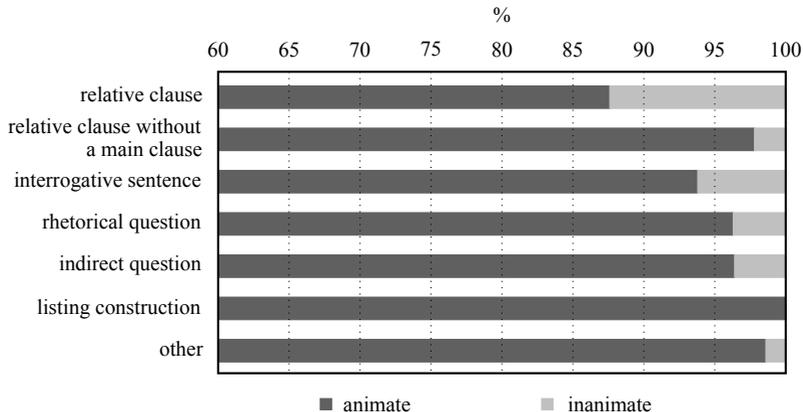
see oli kellega linad pehmes peksetti
 it be:PST:3SG who:COM flax:PL soft:TRL hit:IPS:PST

‘This was [the one] that was used to soften flax.’

Out of all the relative clauses 14.4% were non-headed (341 out of 2,371),

Table 4. The raw frequencies of animate and inanimate referents in different construction types

construction type	animate	inanimate	Σ
relative clause	2,078	293	2,371
relative clause without a main clause	221	5	226
interrogative sentence	76	5	81
rhetoric question	335	13	335
indirect question	160	6	166
listing construction	60	0	60
other	71	1	72
Σ	3,001	323	3,324

**Figure 5.** The percentage of animate and inanimate entities referred to by the pronouns in different construction types

but out of the relative clauses that modified an inanimate entity, only 3.4% were headless (10 out of 293). It is possible that when the pronoun *kes* is used to refer to an inanimate entity, omitting the head of the relative clause from the sentence might cause some misunderstandings and that is why in those cases informants almost always felt the need to clearly state the head in the sentence. Overall, from this analysis we can conclude that out of all construction types, *kes* is most typically used to refer to inanimate entities in headed relative clauses.

3.3 Case marking

To find out how certain grammatical cases are associated with the use of *kes* in regards to referring to inanimate entities, I analysed the cases of all the pronouns in the data.⁹ Eight of the Estonian language's fourteen cases were represented in the data: the nominative, genitive, partitive, elative, allative, adessive, translative and comitative cases. Pronouns in these cases were used to refer to both singular and plural entities (the pronoun *kes* does not itself have plural forms and is always used in the singular), but in this section I have not made a distinction in number. (The relation between the number of the word expressing the referent and the informants' decision to use *kes* to refer to inanimate referents is addressed in § 3.4.) The frequency of the cases of the pronoun *kes* in the data separated by animacy are presented in Table 5.

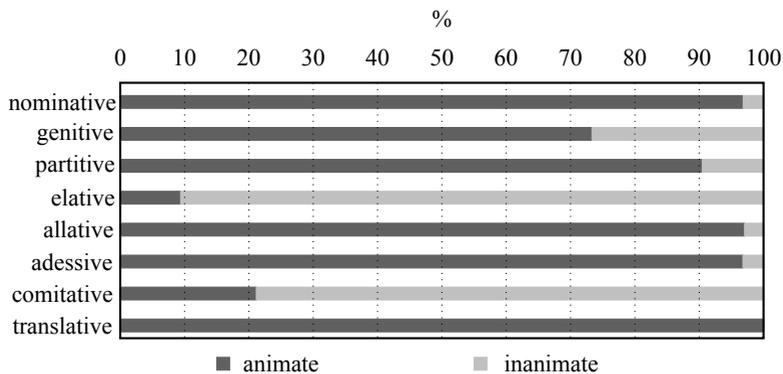
As expected, most of the pronouns in the data were in the nominative case (2,398 pronouns out of 3,324, 72.1%), but using that case to refer to inanimate entities was quite rare: out of all the pronouns in nominative, only 3.2% (76 pronouns) referred to an inanimate entity. The percentage of inanimate entities in all of the cases is represented in Figure 6. It can be seen that every case in the data (except translative) was used to refer both to animate and inanimate entities, meaning that using *kes* to refer to inanimate entities is not limited to certain cases, but is used across the entire paradigm.

Clearly two of the cases stand out from the others: elative, which had 90.6% of all the pronouns referring to an inanimate entity (29 pronouns out of 32), and comitative, which had 78.9% of all the pronouns referring to an inanimate entity (164 out of 208). Note that for every other variable analysed previously, the percentage of animate referents has been dominant, far larger than the percentage of inanimate referents. With the elative and comitative

⁹ For a thorough overview of the Estonian case system, see e. g. Blevins (2008).

Table 5. The frequency of animate and inanimate referents in different cases

case	animate	inanimate	Σ
nominative	2,322	76	2,398
genitive	88	32	120
partitive	57	6	63
elative	3	29	32
allative	33	1	34
adessive	453	15	468
translative	1	0	1
comitative	44	164	208
Σ	3,001	323	3,324

**Figure 6.** The percentage of animate and inanimate entities referred to using the pronoun *kes* in different cases

cases, however, the percentage of the referred to inanimate entities greatly exceeds the number of referred to animate entities, which means that these two cases (or rather the relations they represent) could possibly affect the choice of pronoun the most out of all the variables analysed thus far.

Since elative and comitative differed so much from the other cases, I proceeded to examine their semantic functions in the dataset more thoroughly.

The semantic function of the comitative case is to represent a companion, an instrument, time, manner, condition, etc. (Erelt et al. 2007: 253). In this dataset the pronoun *kes* in the comitative case was only used to refer to inanimate entities when the referent’s function in the sentence was to be an instrument (12–13). The comitative case’s function of being an instrument has developed from the companion’s function through the metaphor AN INSTRUMENT IS A COMPANION (Metslang et al. 2017: 157; see also Stolz et al. 2009). As the instrument may be said to be more grammaticalised than the companion, since the latter usually implies a human participant, whereas the former does not (Heine et al. 1991: 157), a word in the function of an instrument can therefore expectedly be referred to with the animate pronoun in the comitative case.

(12) Mid dialect (Keila)

sealt pealt sai vett võtta ja kellega pesu
 from_there on get:PST:3SG water:PRT take:INF yes who:COM laundry:PRT
pesta
 wash:INF

‘From there one could take water to wash laundry with.’

(13) Western dialect (Kullamaa)

niukse vankri tegin kellega kellega saap
 that_kind:GEN wagon:GEN make:PST:1SG who:COM who:COM get:3SG
kiva põllalt ära vädada
 stone:PL:PRT field:ABL away carry:INF

‘[I] made this kind of a wagon that can be used to carry stones off the field.’

The elative case has many functions in a sentence: it can signify a source, starting time, starting state, material, stimulus, etc. (Erelt et al. 2007: 248–249). In Standard Estonian, the prototypical function of elative is a source or a starting time. However, in this dataset elative was mainly used in the function of material (14). Since there were only three sentences in which elative was

in the function of a source (and even these do not contain a location in the traditional sense; see 15), it can be concluded that *kes* is preferred when referring to entities that are concrete and not locations or abstract times.

- (14) Mid dialect (Juuru)

see pidi *nii pieenikke ja ilus* *olema see lõng*
 it have_to:PST:3SG so fine and beautiful be:SUP that thread
kellest kootti
 who:ELA knit:IPS:PST

‘It had to be so fine and beautiful, the thread that was used to knit.’

- (15) Mid dialect (Kose)

ta old *ikke üks nõu* *kõrvad kõrvad külges kellest*
 it be:PST:PTCL PART one container handle:PL handle:PL attached who:ELA
sa tõstsid teda
 you lift:PST:2SG he:PRT

‘It was a container with attached handles that you lifted it with.’

In fact, when all the inanimate entities in the data referred to by *kes* (altogether 323) were categorised as abstract or concrete, only 13 pronouns referred to an abstract inanimate entity (16–17), while 95% of all the pronouns referred to a concrete inanimate referent. This shows that the use of the pronoun *kes* has not expanded into the pronoun *mis*’s function of referring to abstract things.

- (16) Western dialect (Kullamaa)

põlluarimisest või või kellest see akkab
 farming:ELA or or who:ELA it start:3SG

‘From farming or what it starts from.’

- (17) Mid dialect (Põltsamaa)

luubainaea oli *kiss sellas käis*
 nightmare be:PST:3SG who back:INE walk:PST:3SG

‘[It] was the nightmare that haunted (lit. walked on one’s back).’

3.4 Number

In this section I analyse how the number of the referent is associated with the choice to use the pronoun *kes* to refer to inanimates. To do this, I added the

Table 6. The frequency of animate and inanimate referents separated by the number of the referent

number	animate	inanimate	Σ
singular	2,178	199	2,377
plural	823	124	947
Σ	3,001	323	3,324

number of the word expressing the referred to entity to each sentence. It should be noted that when the referent was only implied in the sentence, but no word for it was included (mostly in the case of questions and indirect questions), the number was always marked as singular if there were no clues in the sentence indicating otherwise.

I hypothesised that since the choice between *kes* ‘who’ and *mis* ‘what’ is in part determined by individuation scale (an individual > a distributive group > a collective), where the likelihood of using *kes* decreases and the likelihood of using *mis* increases from left to right, there should be a higher percentage of plural referents among inanimate referents than among animate referents, since in Standard Estonian the right side of the scale is usually covered by the pronoun *mis* and while referring to inanimates, *kes* is used in the function of *mis*.

The results of the analysis are presented in Table 6. It can be seen that 27.4% of animate referents are in plural, while among inanimate referents the percentage is much higher: 38.4%. The Chi-Squared test finds a significant association between the animacy and the number of the referent ($\chi^2 = 16.68$, $df = 1$, $p = 0.000044$), although according to Cramér’s V ($V = 0.07$) the association between these two variables is not strong at all. Nevertheless, the results support the established hypothesis.

3.5 The interactions of the variables

In this section I conduct a conditional inference tree analysis for my data. The aim of this analysis is to figure out how the variables that I previously analysed separately are associated with the animacy of the referent in relation to each other. This method shows which of the variables influence the choice of using *kes* most significantly and how the variables interact with each other.

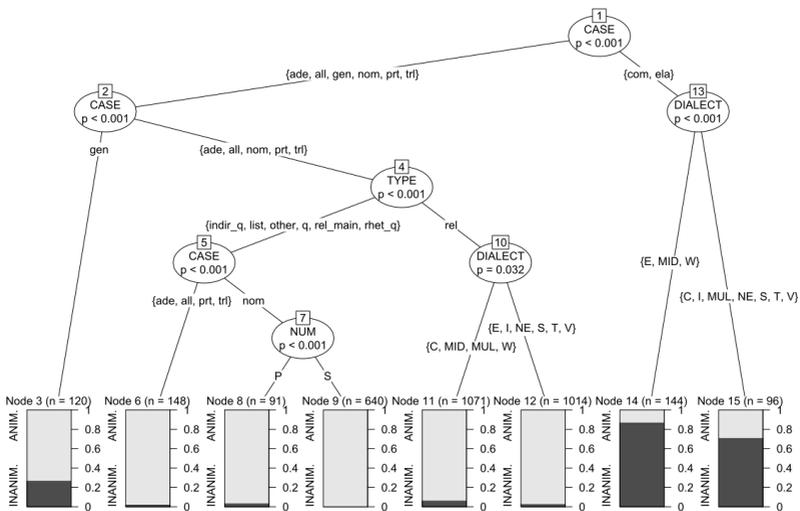


Figure 7. Conditional inference tree for the animacy of the entity that *kes* is referring to

Figure 7 displays the conditional inference tree graph for the animacy of the entity that *kes* is referring to. The variables in this model are the dialect, the construction type, the case and the number of the word expressing the referent. The splits are made in each node by the variable that is most strongly associated with the response: the animacy of the referent. The figure presents all the possible splits significant at the level of 0.05. The bar plots at the bottom show the proportions of animate (light grey) and inanimate (dark grey) referents in each end node, which contain all observations with the given combination of features.

As can be seen in Figure 7, the variable most associated with the animacy of the referent is the case of the pronoun *kes*, which divides the dataset into two based on whether the pronoun is in the elative or comitative case, or some other case. This is in accordance with the preceding analysis which also separated these two cases from the others. Within the group of elative and comitative, a further split is made by dialect: in the Eastern, Mid and Western

dialects, the possibility of an inanimate referent is higher in these two cases (Node 14) than it is in all the other dialects (Node 15), but the percentage of inanimate referents is still relatively high for both of the dialect groups. This is mostly reflective of previous results, except for the Coastal dialect, which typically does use *kes* to refer to inanimate entities in comitative and elative.

The splitting of variables is more complex on the left branch of the tree for the adessive, allative, genitive, nominative, partitive and translative cases. The next significant division is again made by case, which separates genitive from the rest (Node 3), as it is used to refer to inanimates more than the other cases (recall that genitive pronouns referred to the third highest number of inanimate referents).

The sorting of adessive, allative, nominative, partitive and translative on the graph is made by a division of construction types: relative clauses are separated from the rest of the constructions. Relative clauses are divided again by dialect: the Coastal, Mid, Mulgi and Western dialects have a slightly higher percentage of the pronoun *kes* referring to an inanimate entity in relative clauses (Node 11) than the rest of the dialects do (Node 12), although the percentages of inanimate entities are low for both of these groups.

The other construction types (all but the relative clause) are again divided by case, separating nominative from adessive, allative, partitive and translative (Node 6). For the pronouns in nominative the final split is made by the number of the referent: it is possible (although rarely) to refer to an inanimate referent, if the word expressing that referent is in plural (Node 8), but there are no words expressing inanimate referents that are in singular and nominative (Node 9).

The classification accuracy of this tree is 0.95 and the C-index of concordance is 0.89, which both indicate a very good fit.

This analysis shows that the variables included in the method have quite complex interactions with each other. Most of the results correspond with what was discovered before – referring to inanimate entities is most likely to occur in elative and comitative, somewhat likely to occur in genitive, and most prevalent in the Eastern, Mid and Western dialects, although the variable of the dialect is not always significant in affecting the choice to use the pronoun *kes* to refer to inanimate entities. It also showed that for elative, comitative and genitive the construction type is not a significant divider: in those cases, *kes* is used to refer to inanimates in all of the construction types.

Table 7. The number of subdialects, animate and inanimate referents separated by dialect in the new subset

dialect	subdialects	animate	inanimate	Σ
Eastern	3	73	23	96
Mid	15	236	74	310
Western	11	280	74	354
Coastal	3	120	36	156
Insular	3	51	28	79
Σ	35	760	235	995

4 A subset of headed relative clauses

Since the majority of the constructions in the dataset were relative clauses (2,375, 71.4% of all the constructions), I analysed them separately to see whether there are any other variables that affect the use of *kes*. To do that, I made a subset of all the headed relative clauses in the data.

To obtain adequate results about the difference between the use of *kes* in regards to referring to animates and inanimates, I limited this subset to only those dialects that (according to the previous analysis) typically use *kes* to refer to inanimate entities, and of those dialects only the subdialects that had at least 8% of the pronouns referring to inanimate entities. This subset still included sentences with both animate and inanimate entities, but only in those subdialects that had a significant variation in the use of *kes*. The characteristics of this new dataset are described in Table 7.

4.1 Pre- and postnominal relative clauses

Firstly, I intended to find out whether the position of the relative clause in relation to the main clause is connected with the referred to entity's animacy. To that end, I divided all the headed relative clauses according to whether they were preceding (prenominal; 18) or following (postnominal; 19) the main clause (for more about relative clause types, see e. g. De Vries 2001). In the Estonian language the postnominal relative clauses are typically more frequent (Erelt 2004: 407–408), but both options are grammatically correct and both are used in the spoken language.

Table 8. The frequency of prenominal and postnominal relative clauses and referent’s animacy separated by dialects

dialect	prenominal		postnominal		Σ
	animate	inanimate	animate	inanimate	
Eastern	14	0	59	23	96
Mid	71	1	165	73	310
Western	69	1	211	73	354
Coastal	16	1	104	35	156
Insular	20	2	31	26	79
Σ	190	5	570	230	995

(18) Coastal dialect (Jõelähtme)

kes sis kõige kangemb oli et sie sai voidu
 who then most strong:COMP be:PST:3SG that this get:PST:3SG victory:GEN
omale
 for_him/herself:ALL

‘Who was the strongest, that one won.’

(19) Western dialect (Martna)

selle jaoks on niisukke labidas kelle peale paned jahh
 that:GEN for be:3SG this_kind spade who:GEN on:ALL put:2SG yes
ja lükkad ahju
 and push:2SG oven:ILL

‘For that there is this kind of a spade that [you] put [the bread] on and put [it] into the oven.’

The results of the distribution of the relative clauses by position and animacy in the data can be seen in Table 8. Just like in Standard Estonian, the postnominal relative clauses are also prevalent in the dialects. However, for relative clauses that modify a word expressing inanimate entity, the postnominal relative clause seems to be almost the only option used. Only five of the inanimate heads precede a relative clause, while 25% of the animate heads precede a relative clause.

One explanation for this could be, once again, the grammaticalisation of the interrogative marker. If the pronoun *kes* is right next to the word expressing

the referred to entity, then its purpose in the sentence is not to add any semantic value, but just to fill the position of the relative marker to form the sentence. Therefore, the closer the pronoun is to the head, the less important its contrast between animate and inanimate is and it is not always necessary to use a pronoun corresponding to the animacy of the referent. However, when the relative pronoun is further from the head, like it is in the case of the prenominal relative clause, it needs to be more semantically connected in order for the listener to understand which entity the pronoun is referring to (see also Heine & Kuteva 2006: 60–61).

4.2 Referential distance

In this section I look at referential distance between the pronoun *kes* and the word expressing the referred to entity in the sentence. This is why this subset of relative clauses included only headed relative clauses: to look at referential distance at all, a sentence has to have both the pronoun and the head noun that the relative clause is modifying. The goal is to see whether this distance has any effect on choice of the pronoun.

The method for implementing referential distance in the analysis of topic continuity was devised by Givón (1983: 11–13), whose method uses this measurement to assess the difficulty of identifying the referent: the shorter the distance is between references to the same participant, the easier it is to identify that participant. Givón's method measures the distance between participants in clauses, one clause being the minimal possible value (in which case the participant is referred to in the previous clause). This referential distance has been found to be significant in other studies about Estonian dialects and spoken Estonian (see e. g. Kalmus 2009; Lindström et al. 2009; Lindström & Vihman 2017).

In this study, however, I measure the referential distance between the pronoun and the head noun in words instead of clauses: for each relative clause the distance in words was counted between the head of the relative clause and the pronoun *kes* (example 19 with the referential distance of zero, example 20 with a referential distance of five). Note that for prenominal relative clauses, the counting started at the end of the relative clause, i. e. the words in the relative clause were not included in the count. The frequencies of the distances are presented in Table 9.

Table 9. The frequency of the distance (in words) between the pronoun *kes* and the head of the relative clause, separated by animacy, including the percentage of the referential distances respectively for animate and inanimate entities.

distance	animate		inanimate		Σ
	n	%	n	%	
0	430	56.6	97	41.3	527
1	156	20.5	64	27.2	220
2	80	10.5	39	16.6	119
3	38	5.0	22	9.4	60
4	23	3.0	5	2.1	28
5	13	1.7	3	1.3	16
6	11	1.5	3	1.3	14
7+	9	1.2	2	0.9	11
Σ	760	100.0	235	100.0	995

(20) Mid dialect (Ambla)

aga tūdi suri tänavu tall ära nüid kiss teda ikke
 but aunt die:PST:3SG this_year s/he:ADE off now who s/he:PRT always
rohkem sis uolitses ja kasvattas
 more then take_care:PST:3SG and raise:PST:3SG

‘But his aunt died this year who used to take care of him and raise him.’

Table 9 shows that for both animate and inanimate entities the preferred position of the pronoun is still as close as possible to the head of the relative clause. However, for inanimate entities the distance is distributed more equally and there are more pronouns that are at a one-, two- or three-word distance from the head. Even though the Chi-Squared test confirms that there is a significant association between referential distance and entity’s animacy ($\chi^2 = 23.72$, $df = 7$, $p = 0.0013$), the association between these two variables is not strong enough (Cramér’s $V = 0.15$) to draw any definite conclusions.

4.3 The interactions of the variables in headed relative clauses

In this section I conduct a conditional inference tree analysis for the relative clause subset, to determine how the variables that could be coded only for

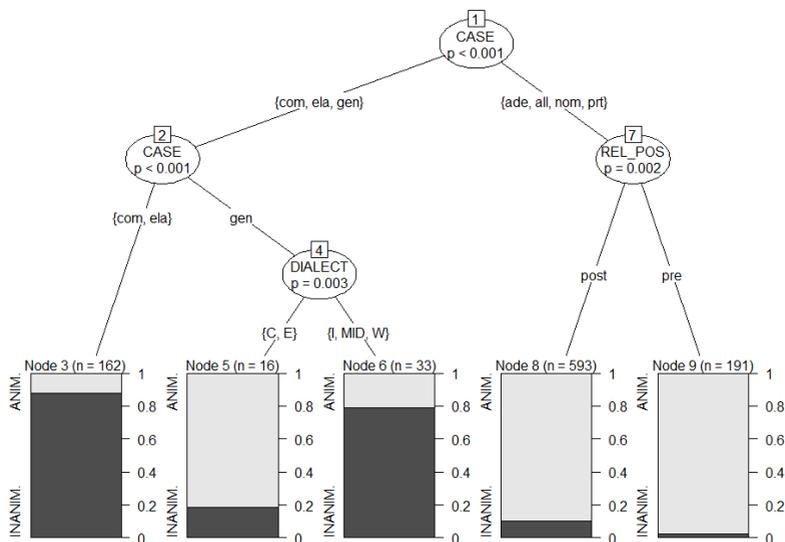


Figure 8. Conditional inference tree for the animacy of the entity that *kes* is referring to in headed relative clauses

headed relative clauses are associated with the animacy of the referent and whether the other variables act similarly to how they did in the conditional inference tree model for the entire dataset.

Figure 8 displays the conditional inference tree graph for the animacy of the entity that *kes* is referring to in headed relative clauses. The variables in this model are the dialect, the case, the number of the word expressing the referent, position of the relative clause relative to the main clause and the referential distance between the pronoun and the head of the relative clause. Again, the splits in each node are made by the variable that is most strongly associated with the response: the animacy of the referent. The bar plots at the bottom show the proportions of animate (light grey) and inanimate (dark grey) referents in each end node, which contain all observations with the given combination of features.

As was the case with the previous conditional inference tree model, the

variable most associated with the animacy of the referent is the case of the pronoun *kes*. The split between the cases is slightly different: elative, genitive and comitative are first grouped together; however, their next split is again made by case and it separates genitive from elative and comitative, which both have a very high percentage of pronouns referring to inanimates (Node 3). The pronouns in genitive are next separated by dialect: *kes* refers to inanimate entities rarely in the Coastal and Eastern dialects (Node 5), but more frequently in the Insular, Mid and Western dialects (Node 6).

For the rest of the cases (adessive, allative, nominative and partitive) the split is made according to the position of the relative clause: the pronoun *kes* is more likely to refer to inanimate entities with a postnominal relative clause (Node 8) than with a prenominal relative clause (Node 9). The percentages for both of these nodes are quite low; nevertheless, it shows that the position of the relative clause only becomes important in those pronoun *kes* cases that are typically not used to refer to inanimates. A closer look at the dataset shows that there are hardly any prenominal relative clauses that have the pronoun *kes* in comitative or genitive (and none in elative), regardless of whether the referred to entity is animate or inanimate. This shows that the use of prenominal relative clauses is not only restricted by the animacy of the referent, but also by the case of the pronoun.

The classification accuracy of this tree is 0.91 and the C-index of concordance is 0.87, which again indicate a very good fit.

As can be seen from the graph, the referential distance between the pronoun and the head of the relative clause (that is, the referred to entity) did not appear to be significant in this model. This confirms my previous conclusion that the distance, while seemingly significantly different for animate and inanimate referents according to the Chi-Squared test, is not actually meaningfully associated with the animacy of the referent or the choice of the pronoun.

The number of the word expressing the referent was also not significant in this conditional tree model, although it proved to be significantly associated with the animacy of the referent in the general analysis of the variable. However, in this subset of the data the animate and inanimate referents are divided very similarly by the number. In fact, for both animate and inanimate referents, 61% of the words expressing the referents were in singular and 39% were in plural, so it is clear that the referent's number is only important in other construction types, but not in headed relative clauses.

5 Discussion and conclusions

In this paper I have examined the way the pronoun *kes* is used in Estonian dialects. While there is a clear distinction between the pronouns *kes* ‘who’ and *mis* ‘what’ in Standard Estonian, in Estonian dialects it is not always so: *kes* can sometimes refer to inanimate entities and *mis* to animate entities. The aim of this paper was to determine which Estonian dialects typically use the pronoun *kes* to refer to inanimate entities and which variables most affect this use.

The first part of the analysis showed that inanimate entities were typically referred to with *kes* in the entire North-Estonian dialect group (except on the island of Saaremaa) and in the Coastal dialect. Overall, however, it was not possible to categorise the dialects and subdialects in a way they have been usually divided in previous studies based on other features. Although southern dialects could be excluded from the typical area of use, in the other dialects there was no clear area where *kes* was used to refer to inanimates: there were isolated subdialect groups all over Northern Estonia where the feature was frequently used. Quite a few dialects and subdialects that are otherwise similar (or even overlapping) differed significantly in the case of this feature (e. g. the Coastal and Northeastern dialects). Nonetheless, it is probable that this use of *kes* has been still typically spread all over Northern Estonia, but is just not always represented in the corpus data used.

Out of all construction types, relative clauses included the highest percentage of inanimate entities that were referred to by the pronoun *kes*, while there were hardly any instances of this with other construction types. I speculate this is due to the grammaticalisation of the interrogative marker, which in the position of introducing (headed) relative clauses is in its most grammaticalised form and has therefore lost some of its distinction in animacy.

When referring to inanimates, the case of the pronoun was mostly elative and comitative, at times also genitive, but the use of other cases was rare. The conditional inference tree analysis showed that while the different variables had complex interactions with each other, the choice to use the pronoun *kes* to refer to inanimate referents was mainly associated with the case of the pronoun (or rather the relations those cases represented). There was also a higher percentage of plural words expressing the referents among inanimate referents than among animate referents.

The second part of the analysis, which included only headed relative

clauses, showed that in the case of inanimate entities only postnominal relative clauses were used, while relative clauses that modified a word expressing an animate entity were both pre- and postnominal. This could, once again, be due to the grammaticalisation of the interrogative marker, whose purpose in postnominal relative clauses (where the pronoun immediately or closely follows the word expressing the referred entity) is not to add any semantic value, but to just fill the position of the relative marker, and therefore its clear distinction in animacy is not at all as relevant as it would be in prenominal relative clauses.

However, the conditional inference tree showed that the position of the relative clause was only significant for those case markings that were typically not used to refer to inanimates. It also appeared that the referential distance between the pronoun and the word expressing the referred to entity was not significantly associated with the choice of the pronoun or the animacy of the referent.

This all shows that although the pronoun *kes* is mainly used to refer to animate entities, there is still significant variation in its use and animacy as a category is not as semantically significant in the dialects as it is in Standard Estonian. It is thanks to corpus-based analysis (which is rather new in the study of dialects) that we are able to examine this (and other) features systematically: prior to this study, there were only scattered notes about the pronoun *kes* being used to refer to inanimate referents, but now it is possible to more clearly define the extent of this feature both in terms of dialect area and sentence context.

In the long term, corpus-based analysis could also help to broaden our understanding of the effects of language contact, which could explain some of the pronoun *kes*'s variation (as several of the contact and cognate languages use interrogative-relative pronouns independently of the referent's animacy). Currently this would be difficult to analyse, due to the fact that the specific features of the local varieties of other languages in Estonia at that time are unknown – we mostly only know the usage of the standard versions of those languages.

Nevertheless it is clear that analysing the counterpart of the pronoun *kes* – the pronoun *mis* ‘what’ – is necessary. Previous dialect overviews (e. g. Kask & Palmeos 1985: 60; Lonn & Niit 2002: 55, etc.) have claimed that *mis* can also refer to animate entities (as opposed to inanimate entities like in Standard Estonian) and the data of this paper shows that *kes* and *mis* are often used to refer to the same entity by the same informant, either in a self-repair clause

or in consecutive finite clauses. Therefore, further analysis would allow us to get a more comprehensive picture of interrogative-relative pronoun variation based on the referent's animacy.

Acknowledgements

This study has been supported by the Centre of Excellence in Estonian Studies (European Union, European Regional Development Fund).

Abbreviations

1, 2, 3	first, second, third person
ABL	ablative
ADE	adessive
ALL	allative
CLI	clitic
CMP	comparative
CNG	connegative
COM	comitative
ELA	elative
GEN	genitive
ILL	illative
INE	inessive
INF	infinitive
IPS	impersonal voice
PART	particle
PL	plural
PRT	partitive
PST	past tense
PTCL	participle
SG	singular
SUP	supine
TRL	translative

References

- Alvre, Paul. 1987. *Soome-ugri keelte ajalooline grammatika I: Sõnaliigid: Arvu- ja käändekategooria* [Historical grammar of Finno-Ugric languages I: Parts of speech: The categories of number and case]. 2nd printing. Tartu: Tartu Riiklik Ülikool.
- Blevins, James P. 2008. Declension classes in Estonian. *Linguistica Uralica* 44(4). 241–267.
- Comrie, Bernard. 1989. *Language universals and linguistic typology: Syntax and morphology*. Oxford: Blackwell.
- Croft, William. 1990. *Typology and universals*. Cambridge: Cambridge University Press.
- Dahl, Östen & Fraurud, Kari. 1996. Animacy in grammar and discourse. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility*, 47–64. Amsterdam: John Benjamins.
- De Vries, Mark. 2001. Patterns of relative clauses. *Linguistics in the Netherlands* 18(1). 231–243.
- Dixon, Robert M. W. 1979. Ergativity. *Language* 55(1). 59–138.
- Erelt, Mati. 1996. Relative words in Estonian relative clauses. *Tartu Ülikooli eesti keele õppetooli toimetised* 4. 9–23.
- 2004. Märkmeid eesti keele komplekslause kohta [Some notes on the Estonian complex sentence]. *Keel ja Kirjandus* 6/2004. 401–413.
- 2014. *Eesti keele lauseõpetus: Komplekslause* [Estonian syntax: Complex sentence]. Tartu: Tartu Ülikooli eesti keele osakond.
- 2017. Liitlause [Complex sentence]. In Erelt, Mati & Metslang, Helle (eds.), *Eesti keele süntaks* [Syntax of the Estonian language], 647–755. (Eesti Keele Varamu III). Tartu: Tartu Ülikooli Kirjastus.
- Erelt, Mati & Erelt, Tiiu & Ross, Kristiina. 2007. *Eesti keele käsiraamat* [Handbook of Estonian]. 3rd, improved edn. Tallinn: Eesti Keele Sihtasutus.
- Fowler, Roger. 1977. *Linguistics and novel*. London: Routledge.
- Givón, Talmy. 1983. Introduction. In Givón, Talmy (ed.), *Topic continuity in discourse: A quantitative cross-language study*, 5–41. (Typological Studies in Language 3). Amsterdam: John Benjamins.
- Heine, Bernd & Claudi, Ulrike & Hünemeyer, Friederike. 1991. *Grammaticalization: A conceptual framework*. University of Chicago Press.
- Heine, Bernd & Kuteva, Tania. 2005. *Language contact and grammatical change*. Cambridge University Press.
- 2006. *The changing languages of Europe*. Oxford: Oxford University Press.
- Hint, Helen & Nahkola, Tiina & Pajusalu, Renate. 2017. With or without articles? A comparison of article-like determiners in Estonian and Finnish. *Lähivõrdlusi. Lähivertailuja* 27. 65–106.

- Hothorn, Torsten & Hornik, Kurt & Zeileis, Achim. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674.
- Juhkam, Evi. 2012. *Harju-Madise murrak* [The subdialect of Harju-Madise]. Ed. by Mari-Liis Kalvik & Helmi Neetar. Tallinn: Eesti Keele Sihtasutus.
- Kalmus, Mervi. 2009. *Isikuihildumine eesti murretes* [Person agreement in Estonian dialects]. Tartu: Tartu Ülikool. (Master's thesis).
- Kask, Arnold & Palmeos, Paula. 1985. *Eesti keele grammatika 2: Morfologia: Arvja asesõnad* [Estonian grammar 2: Morphology: Numerals and pronouns]. 3rd printing. Tartu: Tartu Riiklik Ülikool.
- Kittilä, Seppo & Västi, Katja & Ylikoski, Jussi. 2011. Introduction to case, animacy and semantic roles. In Kittilä, Seppo & Västi, Katja & Ylikoski, Jussi (eds.), *Case, animacy and semantic roles*, 1–26. (Typological Studies in Language 99). Amsterdam: John Benjamins.
- Klavan, Jane & Pilvik, Maarja-Liisa & Uiboed, Kristel. 2015. The use of multivariate statistical classification models for predicting constructional choice in spoken, non-standard varieties of Estonian. *SKY Journal of Linguistics* 28. 187–224.
- Köpp, Kadri. 2001. *Interrogatiiv-relatiivpronoomenid eesti vanemas kirjakeeles* [Interrogative-relative pronouns in old literary Estonian]. Tartu: Tartu Ülikool. (Bachelor's thesis).
- Langacker, Ronald W. 1991. *Foundations of cognitive grammar, vol. 2: Descriptive application*. Stanford: Stanford University Press.
- Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins.
- Lindström, Liina. 2015. *Ülevaade eesti murrete korpusest* [Overview of the Corpus of Estonian Dialects].
(https://www.keel.ut.ee/sites/default/files/www_ut/emk_teejuht2015.pdf)
(Accessed 2020-04-16).
- Lindström, Liina & Kalmus, Mervi & Klaus, Anneliis & Bakhoff, Liisi & Pajusalu, Karl. 2009. Ainsuse 1. isikule viitamine eesti murretes [The first person singular reference in Estonian dialects]. *Emakeele Seltsi aastaraamat* 54 (2008). 159–185.
- Lindström, Liina & Pilvik, Maarja-Liisa. 2018. Korpuspõhine kvantitatiivne dialektoloogia [Corpus-based quantitative dialectology]. *Keel ja Kirjandus* 8–9/2018. 643–662.
- Lindström, Liina & Pilvik, Maarja-Liisa & Plado, Helen. 2018. Nimetamiskonstruktsioonid eesti murretes: Murdeerinevused või suuline süntaks? [Naming constructions in Estonian dialects: Dialectal differences or spoken syntax?] *Mäetagused* 70. 91–126.
- Lindström, Liina & Pilvik, Maarja-Liisa & Ruutma, Mirjam & Uiboed, Kristel. 2015. Mineviku liitaegade kasutusest eesti murretes keelekontaktide valguses [The

- use of the compound past tenses in light of language contacts]. *Võro Instituudi toimõtisõq* 29. 39–70.
- Lindström, Liina & Uiboaed, Kristel. 2017. Syntactic variation in 'need'-constructions in Estonian dialects. *Nordic Journal of Linguistics* 40(3). 313–349.
- Lindström, Liina & Uiboaed, Kristel & Vihman, Virve-Anneli. 2014. Varieerumine tarvis-/vaja-konstruksioonides keelekontaktide valguses [Variation in 'need'-constructions in light of language contacts]. *Keel ja Kirjandus* 8–9/2014. 609–630.
- Lindström, Liina & Vihman, Virve-Anneli. 2017. Who needs it? Variation in experiencer marking in Estonian 'need'-constructions. *Journal of Linguistics* 53(4). 789–822.
- Lonn, Varje & Niit, Ellen. 2002. *Saarte murde tekstid* [Insular dialect texts]. (Eesti Murded VII). Tallinn: Eesti Keele Instituut.
- Metslang, Helle & Habicht, Külli & Hennoste, Tiit & Jürine, Anni & Laanesoo, Kirsi & Ogren, David. 2017. Artikkel õnnesoovidega: Komitatiivi funktsioonidest eri aegade ja registrite eesti kirjakeeles [An article with felicitations: The functions of comitative in literary Estonian of different times and registers]. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 8(1). 149–178.
- Niklus, Lea. 1957. *Otepää murrak* [The subdialect of Otepää]. Tartu: Tartu Riiklik Ülikool. (Diploma thesis).
- Pajusalu, Karl & Hennoste, Tiit & Niit, Ellen & Päll, Peeter & Viikberg, Jüri. 2009. *Eesti mured ja kohanimed* [Estonian dialects and place names]. Ed. by Mari-Liis Kalvik & Tiit Hennoste. 2nd, improved edn. Tallinn: Eesti Keele Sihtasutus.
- Pajusalu, Renate. 1997. Is there an article in (spoken) Estonian? In Erelt, Mati (ed.), *Estonian: Typological studies II*, 146–177. Tartu: Tartu Ülikooli Kirjastus.
- 2005. Anaphoric pronouns in spoken Estonian. In Laury, Ritva (ed.), *Minimal reference: The use of pronouns in Finnish and Estonian discourse*, 107–134. Helsinki: Suomalaisen Kirjallisuuden Seura.
- 2009. Pronouns and reference in Estonian. *Sprachtypologie und Universalienforschung* 62(1–2). 122–139.
- Pook, Hanna. 2018. *Pronoomeni kes kasutusest eesti murretes* [The use of the pronoun *kes* 'who' in Estonian dialects]. Tartu: Tartu Ülikool. (Master's thesis).
- Pung, Lea. 1968. *Tartu-Maarja murrak* [The subdialect of Tartu-Maarja]. Tartu: Tartu Riiklik Ülikool. (Diploma thesis).
- R Core Team. 2018. *R: A language and environment for statistical computing*. (<https://www.R-project.org/>) (Accessed 2020-04-16).
- Ruutma, Mirjam. 2016. *Kaassõnad eesti murretes* [Adpositions in Estonian dialects]. Tartu: Tartu Ülikool. (Master's thesis).
- Ruutma, Mirjam & Kyröläinen, Aki-Juhani & Pilvik, Maarja-Liisa & Uiboaed, Kristel. 2016. Ambipositsioonide morfosüntaktilise varieerumise kirjeldusi

- kvantitatiivsete profiilide abil [Descriptions of the morphosyntactic variation of ambipositions by means of quantitative profiles]. *Keel ja Kirjandus* 2/2016. 92–113.
- Silverstein, Michael. 1976. Hierarchy of features and ergativity. In Dixon, Robert M. W. (ed.), *Grammatical categories in Australian languages*, 112–171. Canberra: Australian National University.
- Stolz, Thomas & Stroh, Cornelia & Urdze, Aina. 2009. Varieties of comitative. In Malchukov, Andrej L. & Spencer, Andrew (eds.), *The Oxford handbook of case*, 601–608. New York: Oxford University Press.
- Szmrecsanyi, Benedikt & Anderwald, Lieselotte. 2018. Corpus-based approaches to dialect study. In Boberg, Charles & Nerbonne, John & Watt, Dominic (eds.), *The handbook of dialectology*, 300–313. Hoboken: John Wiley & Sons.
- Tuldava, Juhan. 1993. *Rootsi keele grammatika* [Swedish grammar]. 4th printing. Tartu: Tartu Ülikool.
- Uiboed, Kristel. 2013. *Verbiühendid eesti murretes* [Verbal constructions in Estonian dialects]. Tartu: Tartu Ülikool. (Doctoral dissertation).
- Uiboed, Kristel & Hasselblatt, Cornelius & Lindström, Liina & Muischnek, Kadri & Nerbonne, John. 2013. Variation of verbal constructions in Estonian dialects. *Literary and Linguistic Computing* 28(1). 42–62.
- Whaley, Lindsay J. 1996. *Introduction to typology: The unity and diversity of language*. Thousand Oaks: Sage.
- Wiedemann, Ferdinand Johann. 2011. *Eesti keele grammatika* [Estonian grammar]. Ed. by Ellen Niit. Trans. by Heli Laanekask. Tallinn: Eesti Teaduste Akadeemia Emakeele Selts. (Original German edition published in 1875).
- Yamamoto, Mutsumi. 1999. *Animacy and reference: A cognitive approach to corpus linguistics*. Amsterdam: John Benjamins.

Contact information:

Hanna Pook
 University of Tartu
 Institute of Estonian and General Linguistics
 Jakobi 2
 51014 Tartu
 Estonia
 e-mail: hanna(dot)pook(at)gmail(dot)com

“Something bad can now happen to me here”: Meaning components of emotion words

Ulla Vanhatalo
University of Helsinki

Heli Tissari
Stockholm University

Taru Lilja
Häme University of Applied Sciences

Kimmo Vehkalahti
University of Helsinki

Mari Siirainen
University of Helsinki

Abstract

This paper reports on how people connect explications of emotion words to the terms they are meant to explicate. We focused on the Finnish counterparts to the following words: *anger*, *disgust*, *fear*, *joy*, *love*, *sadness*, and *surprise*. Our primary findings show that our participants, who were native speakers of Finnish, made the expected matches between Natural Semantic Metalanguage-based explications and the corresponding emotion words. However, there were significant differences between the emotion words, with the match rate ranging from 93% for ‘love’ to 51% for ‘sadness’. This research also contributes to our understanding of the meaning components of emotion concepts, and it may help people to talk about emotions in depth without using the conventional vocabulary for emotions.

Keywords: emotion words, Finnish, meaning components, Natural Semantic Metalanguage (NSM), survey

1 Introduction: The need for empirical testing of NSM explications of emotion words

Talking about emotions is a quotidian and ubiquitous action. People talk about emotions such as ‘anger’, ‘joy’, ‘sadness’, ‘fear’, ‘surprise’, and ‘love’ in

everyday language. These concepts are also taught to children and young people in schools and in official materials that support recognizing and managing emotions (e. g. The Finnish Association for Mental Health 2019). Emotion words are labels for these concepts, and they guide our common conceptions of emotions. However, their meaning may vary from person to person. A more analytical and simple way of defining and describing emotions could be useful. Thus, in our study we sought to test how well so-called Natural Semantic Metalanguage explications really describe and define emotion words in Finnish.

For 40 years or so, proponents of Natural Semantic Metalanguage (NSM) have claimed that concepts of emotions can be explicated by simple words (Wierzbicka 1999; Harkins & Wierzbicka 2001; Goddard 2002). Yet, there has been little empirical proof for these claims. Regardless of what at first glance appears to be its theoretical and linguistic nature, the question has a strong practical impact on many fields outside of linguistics. Rapidly gaining ground is the argument that simple language can discuss issues as complex as emotions. In this paper, we test whether laypeople really understand the meaning components suggested by linguists. We approach the question by investigating how well native speakers of Finnish can match emotion explications with the term they are meant to explicate and identify the components of meaning.

The explications and meaning components used in this study are based on Natural Semantic Metalanguage (NSM). NSM is a mini-language consisting of about 65 simple words, the so-called semantic primes. According to NSM theory, these primes are universal concepts and can be found in all languages of the world. With these 65 words, all the remaining words in any language can be explained. This linguistic theory, originally proposed by Anna Wierzbicka and further developed by Cliff Goddard and others, combines structural and cognitive aspects of word meaning (e. g. Wierzbicka 1992a; Goddard 2008). Lexical meaning is seen as being both deconstructable and explicable. Table 1 lists the semantic primes in English, showing how they correspond to simple concepts that people can be expected to be familiar with without further explanation.

Emotion vocabulary has been the focus of many NSM studies over the last several decades, and a wide selection of explanations has been published, covering various languages and emotion concepts (e. g. Wierzbicka 1999; Harkins & Wierzbicka 2001). One of the revolutionary notions in NSM emotion studies is related to categorizing emotions: most emotion

Table 1. Semantic primes (English exponents; Goddard & Wierzbicka 2014). Exponents of primes can be polysemous (i. e. they can have other additional meanings). Exponents of primes may be words, bound morphemes, or phrasemes. They can be formally, that is, morphologically, complex. They can have combinatorial variants or allolexes (indicated with ~). Each prime has well-specified syntactic (combinatorial) properties.

substantives:	I, YOU, SOMEONE, SOMETHING~THING, PEOPLE
relational substantives:	BODY KINDS, PARTS
determiners:	THIS, THE SAME, OTHER~ELSE
quantifiers:	ONE, TWO, SOME, ALL, MUCH~MANY, LITTLE~FEW
evaluators:	GOOD, BAD
descriptors:	BIG, SMALL
mental predicates:	KNOW, THINK, WANT, DON'T WANT, FEEL, SEE, HEAR
speech:	SAY, WORDS, TRUE
actions, events, movement:	DO, HAPPEN, MOVE
location, existence, specification:	BE (SOMEWHERE), THERE IS, BE (SOMEONE/SOMETHING)
possession:	(IS) MINE
life and death:	LIVE, DIE
time:	WHEN~TIME, NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME, MOMENT
place:	WHERE~PLACE, HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE, TOUCH
logical concepts:	NOT, MAYBE, CAN, BECAUSE, IF
augmentor, intensifier:	VERY, MORE
similarity:	LIKE

explications include either a “feeling something good” or “feeling something bad” component (Wierzbicka 1999). Emotion words can thus be divided roughly into two categories, within which they can be further differentiated by using simple, language-independent prime words. Emotion words can be deconstructed into components of meaning; if needed, they can also be compared component by component in different languages. Many (e.g. Harkins 2001) of the published NSM-based emotion word explications are based on linguistic analyses by author introspection and supported by examples from linguistic corpora. The explications typically follow the semantic template specifically constructed for an emotion-based vocabulary, as can be seen in the cases of the English word *happy* or the non-translatable Japanese *amae*:

(1) **someone X is happy (at this time):**

Someone X thinks like this at this time:

“many good things are happening to me as I want
I can do many things now as I want
this is good”

because of this, this someone feels something good at this time
like someone can feel when they think like this
(Goddard & Wierzbicka 2014)

(2) **someone X feels *amae* (towards Y) at this time:**

someone X thinks like this at this time (about someone Y)

“this someone can do good things for me
this someone wants to do good things for me
when I am with this someone, nothing bad can happen to me
I want to be with this someone”

because of this, this someone feels something good at this time
like someone can feel when they think like this
(Wierzbicka 1998)

As can be seen from the above explications, NSM focuses on people’s thoughts about the emotion they experience. Another way to approach emotions would be to focus on how the specific emotion is felt in the body, as in the GRID paradigm. Ye (2013) has compared the NSM and GRID paradigms and suggested that they complement each other.

According to the principles of NSM, explications of words must be simple enough to be understood by laypeople (Goddard 2011: 65). The

necessity of testing the empirical evidence of explications has been discussed, and at least one study on the empirical testing of NSM explications has been carried out. Gladkova et al. (2016) conducted a study on the English interjections *wow*, *gosh*, *gee*, *yikes*, *yuck*, and *ugh*, in which they used a survey technique to collect experimental data ($N = 101$). The purpose was to evaluate a set of NSM explications of interjections. Their results show that the explications of interjections were well accepted by laypeople overall, with the proposed explications receiving positive rankings on a Likert scale of 1–7 (grand mean score = 5.38, range = 4.6–6.2). Gladkova et al. (2016) observed that the differences in the scores for different words could be associated with the clarity of the explications or even with the level of the participants' familiarity with the words in question. They also assessed the effect of background variables (age, gender, [non-]native speaker status, and background knowledge about NSM) on the respondents' evaluations, noting that the background variables had little impact on the results.

Gladkova et al. (2016) made a solid beginning with their systematic empirical testing of NSM explications, yet there is still a need to continue the research. A key question is whether their findings on interjections hold true for other word classes or semantic fields, such as emotion vocabulary. Harkins's study (2001) on words for anger in Central Australia is relevant for a different reason, because she worked with aboriginal speakers to collect, understand and explain words and their nuances. In our study, the informants did not play an equally active role in contributing to how the words should be explicated.

Recently, there has been growing interest in the use of NSM-based applications outside of linguistics and outside of academia (e. g. Goddard 2018; Vanhatalo & Torkki 2018). This kind of non-academic usage shows the practical value of the NSM approach. The NSM primes themselves seem to be intuitively accepted by ordinary people. But are the NSM explications or some parts of them accepted as well? Could the meaning components of emotion words be used in psychological practice, for instance?

Another aspect to consider is the translatability of the NSM explications and the emotion concepts. Nearly all of the semantic primes can be found in most languages. The explications themselves, if carefully crafted, should be translatable, at least to some extent. Yet, the conventional vocabulary of emotions is often culturally dependent, and meanings may vary radically from one language to another. However, it appears that this kind of semantic diversity does not apply to all emotion concepts equally. While some concepts

are almost impossible to translate (e. g. German *Angst*, Wierzbicka 1999; Japanese *amae*, Wierzbicka 1998), other concepts seem to be almost the same across languages (e. g. *fear*-like concepts, Wierzbicka 1999).

The ultimate goal of this study is to determine if the explications and meaning components of emotion words, as suggested by NSM-oriented linguists, could be useful for non-linguistic purposes. Our purpose in this particular study is not to convince readers of the selection of the NSM primes; there is already abundant literature on various aspects of NSM, including critiques. Nor do we try to convince readers of the accuracy of the current NSM explications. Rather, we take both the primes and the explications as they are published in previous NSM literature. We agree with some but not all aspects of NSM, and we recognize that there are some controversial issues in the theory. Nevertheless, what we try to do is to increase the mutual discussion between NSM and other approaches (such as linguistic survey studies) to emotion vocabulary. Our strategy here is to expand the number of methodological devices traditionally used in the NSM framework and, by doing so, seek acceptable and valuable parts of NSM explications of emotion concepts.

2 Materials and method

For the current study, we selected six emotion concepts: ‘anger’, ‘disgust’, ‘fear’, ‘joy’, ‘sadness’, and ‘surprise’. We added the concept of ‘love’ to the set. Although contemporary psychological emotion research does not support traditional basic emotion theory any more (Scarantino 2014: 334–376), these six concepts are the same as Ekman’s classic concepts (Ekman later dropped ‘surprise’ from the list). However, our list is not based solely on these “universal emotion concepts” but also on the research findings of Tuovila (2005), who made Finnish speakers write down all the emotion words they could remember. There were twenty words that were mentioned most frequently. These words were (the order is the frequency order) *viha* ‘hatred’, *ilo* ‘joy’, *rakkaus* ‘love’, *suru* ‘sadness’, *pelko* ‘fear’, *onnellisuus* ‘happiness’, *kateus* ‘envy’, *ahdistus* ‘anxiety’, *väsymys* ‘tiredness’, *masennus* ‘depression’, *tuska* ‘agony’, *ihastus* ‘admiration’, *tyytyväisyys* ‘contentment’, *inho* ‘disgust’, *jännitys* ‘excitement’, *pettymys* ‘disappointment’, *kaipa* ‘longing’, *rauhallisuus* ‘tranquility’, *ikävä* ‘longing’, and *toivo* ‘hope’. Here we can see that the emotion terms that come to the minds of Finnish speakers

indeed include: ‘hatred’ (*viha*), ‘joy’ (*ilo*), ‘love’ (*rakkaus*), ‘sadness’ (*suru*), ‘fear’ (*pelko*), and ‘disgust’ (*inho*). The term *hämmästys* ‘surprise’ is missing from this list; we took it from Ekman’s classic list of emotions.

We created an online survey and checked how well our participants, all of whom were native speakers of Finnish, could match NSM-based emotion explications and meaning components with the words they are meant to explicate. We tested both the full explications and the separate parts. Participants matched both the explications and the parts with emotion words, so that they could choose one or more, depending on their preference. They could also suggest an emotion word themselves.

Even though this paper has somewhat similar objectives to those of Gladkova et al. (2016), our setting differs in several ways. Partly based on the suggestions made there, we chose nouns (instead of interjections) and made sure our words were normal standard language (no infrequent or possibly unfamiliar words). We also asked our participants to match explications and parts of explications with the explicated words instead of seeking mutual similarity between words and explications by a Likert scale, as done by Gladkova et al.

2.1 Choosing the emotion words

Seven emotion words were ultimately chosen for this study: *viha* ‘anger, hatred’, *inho* ‘disgust’, *pelko* ‘fear’, *ilo* ‘joy’, *rakkaus* ‘love’, *suru* ‘sadness’ and *yllätys* ‘surprise’.¹ These words are based, on one hand, on Ekman’s well-known facial expression studies (Ekman 1993; 1999a; 1999b) and, on the other hand, on Tuovila’s (2005) studies on basic Finnish emotion terms. The set includes the concept of love, which is included in Tuovila’s list of 20 basic emotion terms in Finnish, although it is left out of Ekman’s list (Ortony & Turner 1990).

There were several aspects to consider as we started to select emotion words for our study. First of all, we had to find words that had been studied and explicated in NSM terms. As there are no established explications available for Finnish emotion vocabulary (Tuovila 2005),² we had to begin with work

¹ In order to make this paper easier for a non-Finnish reader, we use the English translations throughout the text. The original survey was addressed to a Finnish audience and was entirely in the Finnish language. Some translations are provided in the appendices to this paper.

² Tuovila (2005) presents many Finnish NSM explications of Finnish emotion words. Many of these seem to be translated more or less directly from English, yet the translation process has not

that had been done for the English language. Second, we mainly chose words with explications intuitively matching those in Finnish (*disgust, fear, joy, sadness, surprise*), although we were well aware of the mismatch between the English word *anger* and Finnish *viha*, which we have studied separately (Tissari et al. 2019). Third, we wanted to use words with substantially different meanings, as the study was not going to focus on synonyms. Fourth, we wanted to link our study with psychology, and thus we chose emotions that have been the focus of psychological studies.

While separating the explications into individual meaning components, we realized that four out of the six emotion words in the Ekman set were negative (*viha* ‘anger, hatred’, *inho* ‘disgust’, *pelko* ‘fear’, *suru* ‘sadness’), and all have the component of “feeling something bad”. Only one word (*ilo* ‘joy’) had the component of “feeling something good”, while one word (*yllätys* ‘surprise’) was without a component of “feeling something good/bad”. The dichotomy between positive and negative emotions has also been observed and discussed in research on the structure of emotions (Watson & Tellegen 1985; Barrett & Russell 1998). Positive psychology argues that focus on the latter is due to the long tradition of clinical psychology’s emphasis on the negative, diseases and illnesses (Wood & Johnson 2016). Suggestions have been made to shift the discipline of clinical psychology to have an integrated and equally weighted focus on both positive and negative functioning, also in emotion research (Wood & Tarrrier 2010: 819). Instead, there has been surprisingly little concrete change in the conceptualizing of basic emotions, considering the vast amount of lateral theories about emotions that exist. In basic emotion research, the Ekman set has persisted (Ekman 2016). The other commonly used and frequently studied sets of emotions mainly concentrate on the same vocabulary: *fear, sadness, anger* and *joy* are usually included, while *disgust* and *surprise* are often the first to be left out (e. g. Jack et al. 2014).

The larger number of negative emotions should not be a problem for NSM, as one of its crucial methodological features is the power to differentiate between close meanings. However, too many components with negative meaning may have confused our study participants; therefore, in order to balance our study setting, we decided to include a positive emotion – the word

been justified or examined in detail. As in NSM literature based on the English language, here too there are various explications (some made after 2005) for the words studied in this paper. We thus decided to choose our explications independently. The only explication from Tuovila that we chose is ‘love’.

rakkaus ‘love’ – in the set. The explication of the word *rakkaus* ‘love’ in our study is special compared to the other explications. Perhaps a bit surprisingly, it was hard to find published explications of the English word *love* in the NSM literature. We first only found Wierzbicka’s (1999: 293–294) discussion of the Ifaluk concept of *fago*, which can be glossed as ‘love’ or ‘compassion’. The word *love* is challenging to define, of course, because it can be used for different degrees of emotion and loves of different kinds, as in “I love it!”.³ As *rakkaus* ‘love’ is one of the top four emotion words (*viha*, *ilo*, *rakkaus*, *suru*) in Finnish (Tuovila 2005), and there also happens to be a published explication for that word in Finnish (Tuovila 2005), we found *rakkaus* to be the perfect addition to our otherwise fairly negative emotion set. We decided to use Tuovila’s definition, an extended version of the rather simple definition of *love* that Wierzbicka (1992b: 145) presents. We also discussed the possible polysemy or synonymy of the words selected for this study, especially with the concepts of *ilo* ‘happiness, joy’ and *hämmästys / yllätys* ‘surprise’, but did not find that to be a problem.

2.2 Choosing, translating, adjusting, and deconstructing the semantic explications

Transforming the explications from the previous NSM research for our current survey required a process of choosing the most appropriate explications, translating these from English-based NSM to Finnish-based NSM, adjusting some parts to be more readable by laypeople, and, finally, separating the full explications into independent meaning components. In this section, we describe the main steps in the process.

The emotion concepts chosen for this research have been examined in multiple earlier studies: *anger* and *anger*-like or *wut*-like emotions (Wierzbicka 2014: 80–83), *disgust* (Goddard 2014: 79–81), *fear* and *fear*-like emotions (Wierzbicka 2014: 82), *joy* (Wierzbicka 1999: 50), *happy* (Wierzbicka 1999: 52; Goddard 2011: 110), *rakkaus* ‘love’ (Tuovila 2005: 112; Wierzbicka 1999: 293–294), *sadness* (Wierzbicka 1998) and *surprise* (Goddard 2015: 298). Most of the emotion words have several published explications, some of which have been rethought and rewritten by the original author, while other explications have been published by different authors. We compared the published versions and applied our intersubjective

³ We thank an anonymous reviewer for pointing this out.

linguistic intuition in choosing explications from those available. Our aim was to find the versions which sounded the most accurate to us. We did not mix the versions of explications. Instead, we stuck to the one that we found the most convincing. The explication of *rakkaus* ‘love’ was based on emotion studies in the Finnish language and published in Finnish (Tuovila 2005). All other explications involved English emotion words and were published in English. The semantic template of the explication of *rakkaus* ‘love’ was somewhat different from the others, as it included parts not widely used in the NSM literature. These observations may indicate that the explication process of love had been done separately from the NSM community.

All the explications originally published in English-based NSM were collaboratively translated into Finnish-based NSM by the authors of the study. There were no specific challenges in the translation process. All the translations can be found in Appendix A.

Our goal was to make the survey questions short, unambiguous, and easy and quick to answer. In relation to this adjustment process, we separated the full explications into parts, in order to obtain individual meaning components for each word. We provide the original explication of ‘fear’ here as an example:

- (3) **fear**-like emotions (Wierzbicka 2014: 78)

It can be like this:

Someone thinks like this:

“Something bad can happen to me here now

I don’t want this

I don’t want to be here

I want to do something because of this”

When this someone thinks like this, this someone feels something bad because of this, like people often feel when they think like this

After the translation and adaptation process, the meaning components of ‘fear’ looked like this:

- (4) *minulle voi tapahtua täällä nyt jotain pahaa* ‘something bad can now happen to me here
en halua tätä ‘I don’t want this’
en halua olla täällä ‘I don’t want to be here’
tämän takia haluan tehdä jotain ‘I want to do something because of this’
kun ajattelen tätä, minusta tuntuu pahalta ‘when I think about this, I feel bad’

All the original explications and adjusted versions used in our survey can be found in Appendix A. In the adjustment process, we used the following techniques:

- a) As all the words belong to the same semantic field, the outer parts of the semantic templates (mainly the so-called lexico-syntactic frames) were removed as they were unnecessary (see similar processes in Gladkova et al. 2016). For example, we removed parts such as *it can be like this* or *X felt something because X thought something* or *sometimes a person thinks*.
- b) The questions were made personal by changing the point of view from third-person observing to first-person experiencing. This happened partly in the course of removing the lexico-syntactic frames, partly by changing the actual personal pronouns; for instance, the original *when this person thinks this, this person feels something very good* became *kun ajattelen tätä, minusta tuntuu hyvältä* ‘when I think this, I feel good’, or the original *sometimes a person thinks* became *joskus ajattelen näin* ‘sometimes I think like this’.
- c) Some pronouns were replaced with nouns; for example, *haluan hänelle hyvää* ‘I want good for him/her’ was changed to *haluan toiselle ihmiselle hyvää* ‘I want good for another person’, or *I didn’t think before that this someone can do something like this* was changed to *en ollut ennen ajatellut, että tämä ihminen voi tehdä jotain tällaista* ‘I had not previously thought that this person could do something like this’. These changes were necessary, especially for survey questions including individual components.
- d) Finally, we checked the overall reading of all the questions. Minor changes (e. g. involving time or syntax) were made.

The outcome of these processes was a set of 33 questions, including 7 full explications and 26 individual semantic components.⁴ We first tried out the

⁴ An unfortunate error in the questionnaire was revealed during manuscript proofreading. One component of *rakkaus* ‘love’ had accidentally occurred twice in the survey. Two alternative wordings had been used: *haluan hyvää toiselle ihmiselle* and *haluan toiselle ihmiselle hyvää* ‘I want good for another person’. We apologize for our mistake. Fortunately, this error does not seem to have influenced the other emotion words in this study. Interestingly, the error shows over 8% variation in acceptance rates between the two alternative components, see Figures 4 and 5.

questions in a pilot project with 16 participants, at which point further minor adjustments were made.

In the course of this process of making changes, we sought a balance between faithfulness to the original NSM research and the feasibility of the survey setting. It was clear that the original explications as such would not have worked in a survey, but we did not want to change them too much, so as not to lose the possibility of comparison between NSM studies. We made great efforts to determine the minimal changes needed. Even if most of the individual meaning components worked very well, we were – and are – aware of the slightly unnatural tone of some of the components, such as the Finnish wording in the phrase *on olemassa joku ihminen* ‘there is someone / a person’ (a component of *rakkaus* ‘love’). Yet, we did not want to change the words too much, and we wanted to keep to the original explications as much as possible. It should also be emphasized here that no new semantic components were created by our team; instead, all the components came from previous studies.

2.3 The questionnaire and data sampling

2.3.1 The questionnaire

The online questionnaire consisted of 33 questions, including 7 questions with full explanations and 26 questions with individual semantic components. The core question was: “To which emotion is this [expression] linked? Choose one or more options.” In addition to these multiple choice questions, the questionnaire included translation tasks, background questions, and the “consent to participate” portion. Participants saw just one question at a time on the computer screen, as shown in Figure 1.

The core questions were all multiple choice, and the same set of answers was always provided. The choices were 15 emotion words together with “I don’t know” and finally “some other word; specify which”. This meant that the participant was not forced to choose between the 15 emotion words but could suggest his or her own. Among the 15 emotion words, the seven tested words (*viha* ‘anger, hatred’, *inho* ‘disgust’, *pelko* ‘fear’, *ilo* ‘joy’, *rakkaus* ‘love’, *suru* ‘sadness’ and *hämmästyys* ‘surprise’) always appeared. In addition, there were filler words, consisting of *ahdistus* ‘anxiety’, *häpeä* ‘shame’, *kateus* ‘envy’, *onnellisuus* ‘happiness’, *riemu* ‘elation’, *tyytyväisyys* ‘contentment’, and *ylpeys* ‘pride’. Four of these refer to different emotions (*ahdistus* ‘anxiety’, *häpeä* ‘shame’, *kateus* ‘envy’, *ylpeys* ‘pride’), while three

minulle voi tapahtua täällä nyt jotain pahaa
 en halua tätä
 en halua olla täällä
 tämän takia haluan tehdä jotain
 kun ajattelen tätä, minusta tuntuu pahalta

Mihin tunteeseen tämä mielestäsi liittyy? Merkitse rastilla yksi tai useampi vaihtoehto.

ahdistus
 hämmästyks
 häpeä
 ilo
 inho
 kateus
 onnellisuus
 pelko
 rakkaus
 riemu
 suru
 tyytyväisyys
 viha
 ylpeys
 en osaa sanoa
 jokin muu, mikä?

<< Edellinen | Seuraava >>

tiedän, että en voi tehdä mitään

Mihin tunteeseen tämä mielestäsi liittyy? Merkitse rastilla yksi tai useampi vaihtoehto.

ahdistus
 hämmästyks
 häpeä
 ilo
 inho
 kateus
 onnellisuus
 pelko
 rakkaus
 riemu
 suru
 tyytyväisyys
 viha
 ylpeys
 en osaa sanoa
 jokin muu, mikä?

<< Edellinen | Seuraava >>

Figure 1. Two screenshots of the survey form. Translation of the upper part: ‘Something bad can now happen to me here / I don’t want this / I don’t want to be here / I want to do something about this / When I think about this, I feel bad / To which emotion is this statement linked? Choose one or more alternatives’. Translation of the lower part: ‘I know that I cannot do anything / To which emotion is this linked? Choose one or more alternatives’.

had meanings fairly similar to the words studied (*onnellisuus* ‘happiness’, *riemu* ‘elation’, *tyytyväisyys* ‘contentment’). The words were always in the same alphabetical order. What is also important to remember is that the respondents had to make at least one choice, but multiple choices were also possible. This means that the participants may have used different strategies in selecting their answers. Participants could select multiple words, just one, or suggest one that was completely their own.

2.3.2 Participants

The data were gathered as convenience samples from two groups of Finnish university students in supervised situations. The students answered the questions as a part of their coursework in Introduction to Cognitive Psychology and Introduction to General Linguistics. The allotted time to answer the questionnaire was limited to 15–20 minutes, but many participants completed the survey in 10 minutes. The online questionnaire platform that we used did not allow us to deny access to participants to go back to their answers and change them, but in the directions on how to fill out the survey we stressed the importance of participants not returning to their previous answers and making changes. Because the situation was supervised, it was also possible to monitor the completion process. The participants were not allowed to see the survey beforehand. We received 147 answers, out of which 130 were chosen for the final analysis, since 17 respondents had another language than Finnish as their first language. There were 97 female and 30 male respondents (and 3 other). Their ages varied from 18 to 54 years, with a mean age of 26.8 years and a median of 24.0 years. Because previous research on Natural Semantic Metalanguage explications of interjections has shown that respondents’ background variables (age, gender, linguistic study) did not significantly influence the results (Gladkova et al. 2016), we did not analyse our results with respect to these.

3 Findings

Our primary findings answer our research question positively: our participants were often able to match the explications with their intended emotion words. However, there were significant differences in the matches between the emotion words and the explications: the match rate ranged from 93% for *rakkaus* ‘love’ to 51% for *suru* ‘sadness’. Our secondary findings deal with

individual meaning components, which were also matched with the words explicated; here the match rate varied even more.

In this section, we first present the overall picture of the primary findings for the full explications, followed by comments on each emotion word, including observations on the top unexpected matches. Second, we point out some notions about matching individual components with the words explicated. In evaluating the findings, we must emphasize that the participants were allowed to choose more than one answer per question. This means that the same participant may have both matched a particular explication or meaning component with the intended word and matched it with some other word.

3.1 Matching the full explications with the emotion words

Rakkaus ‘love’ and *pelko* ‘fear’ were the two best recognized emotion concepts in our study, *rakkaus* being matched with its explication by 93% of the participants and *pelko* by 76%. *Yllätys* ‘surprise’ and *suru* ‘sadness’ were matched with their explications relatively poorly, with roughly half of the participants recognizing them. The overall findings on matching the emotion words and their full explications are presented in Figure 2. Interestingly, many participants made frequent matches between the explanations given and unexpected words, including the filler words (Figure 3). Remarkably few participants chose the option “I don’t know”. The percentage of informants who did this ranged mostly from 0% (*pelko* ‘fear’, *viha* ‘anger’, *rakkaus* ‘love’, *ilo* ‘joy’) to 2% (*suru* ‘sadness’), with the case of *yllätys* ‘surprise’ being exceptional: 14% of the participants chose “I don’t know”.

Below we take a closer look at each emotion word separately, giving special emphasis to certain findings.

- *rakkaus* ‘love’: The match rate for ‘love’ was 93%, meaning that 121 participants connected ‘love’ with the explication, while 7% (9) participants did not make that connection. However, the participants also connected some other words with this good feeling. 76% (99) suggested the word explicated here could be *onnellisuus* ‘happiness’, while 54% (70) suggested *ilo* ‘joy’ and 44% (57) *tyytyväisyys* ‘contentment’. No one chose “I don’t know” and just 2% (2) provided an answer to the open question.
- *pelko* ‘fear’: 76% of the participants (99) made the expected choice

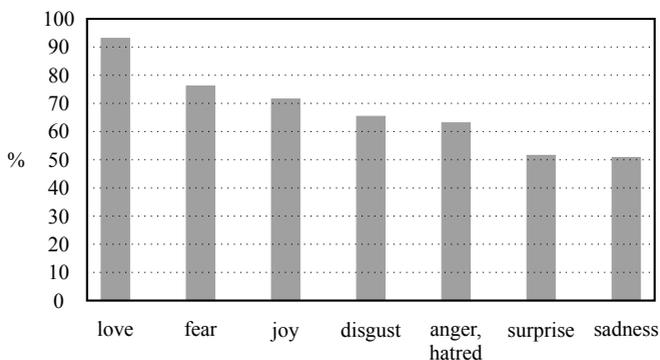


Figure 2. The bars show the percentages of participants ($N = 130$) who made the expected matches between emotion words and their explanations. The participants were allowed to choose more than one answer per question. This means that besides an expected answer, the same participant may also have selected an unexpected answer for some word(s).

by connecting ‘fear’ with the given explanation, while 24% (31) did not make this connection. Curiously, the most common other answer, *ahdistus* ‘anxiety’, got a much higher score than the expected answer, with 92% of the participants (120) choosing it. The next two most common other answers were *inho* ‘disgust’ 32% (41) and *häpeä* ‘shame’ 28% (36). None of the participants chose “I don’t know”, while 4% (5) provided open answers.

- *ilo* ‘joy’: 72% of the participants (93) chose ‘joy’ for the given explanation, while 28% (37) did not make this choice. The top three other answers got scores close to the word in question, as *onnellisuus* ‘happiness’ was chosen by 67% (87), *tyytyväisyys* ‘contentment’ by 55% (71), and *riemu* ‘elation’ by 48% (62) of participants. “I don’t know” was chosen by 0% of participants, while 2% (3) provided answers to the open question.
- *inho* ‘disgust’: 65% of the participants (85) recognized ‘disgust’ as the word explained, and 35% of the participants (45) did not make the connection. The top three other answers were *ahdistus* ‘anxiety’

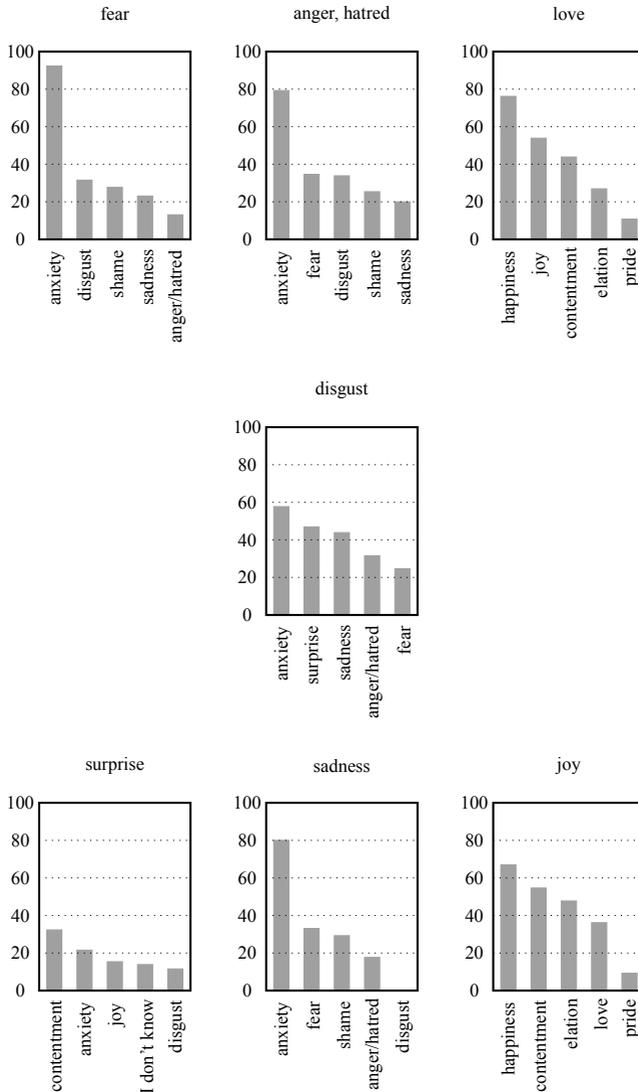


Figure 3. Findings on unexpected answers. The headings give the word actually explicated. The numbers on the left show the percentage of participants ($N = 130$) who chose other words to match the explication. The other words included other emotion words studied in this setting as well as filler words.

58% (75), *yllätys* ‘surprise’ 47% (61), and *suru* ‘sadness’ 44% (57). “I don’t know” was chosen by 1% of the participants (1), and 5% of the participants (6) gave open answers.

- *viha* ‘anger, hatred’: 63% of the participants (82) made the connection to ‘anger, hatred’, while 37% (48) did not. Surprisingly, the most popular unexpected answer got a much higher score than *viha* ‘anger, hatred’ itself, with 79% of the participants (103) connecting *ahdistus* ‘anxiety’ with the explication for *viha* ‘anger, hatred’. Much less popular but still significant in number were the next most frequent unexpected answers: *pelko* ‘fear’, suggested by 36% of the participants (45), and *inho* ‘disgust’, suggested by 32% of the participants (41). “I don’t know” was not chosen at all, and 2% of the participants (3) gave open answers.
- *yllätys* ‘surprise’: Slightly more than half of the participants, 52% (67), connected ‘surprise’ with its explication, with nearly half not making the connection. The top three other answers leave a great deal of room for interpretation, the most popular choices being *tyytyväisyys* ‘contentment’ by 32% of the participants (42), *ahdistus* ‘anxiety’ by 22% of the participants (28), and *ilo* ‘joy’ by 15% of the participants (20). “I don’t know” has the only remarkably high score here, 14% (18), followed by a relatively high score of open answers, 5% (7).
- *suru* ‘sadness’: Roughly half of the participants 51% (66) made the connection between ‘sadness’ and its explication, while almost the same number (64) did not make the connection. Significantly, *ahdistus* ‘anxiety’ stands out as very high, compared to the other answers: 80% of the participants (104) thought that it would be the emotion word connected with the explication for *suru* ‘sadness’. The two next most popular answers were *pelko* ‘fear’, chosen by 33% of the participants (43) and *häpeä* ‘shame’, chosen by 29% of the participants (38). “I don’t know” was chosen by 2% of the participants (2), and open answers were given by 8% (11).

In Figure 3, we show the top five unexpected word choices for each explication. Along with the statistics concerning expected answers presented earlier in this subsection (Figure 2), these statistics provide useful insights into how the participants in this study connected the semantic fields of the

emotion words. The highest percentages, especially for *pelko* ‘fear’, *viha* ‘anger, hatred’, *rakkaus* ‘love’, and *suru* ‘sadness’, indicate that the meaning explications of these words are very close to those for some other emotion words. Our findings suggest that there may be a particularly strong connection between the semantic explications of fear, anger, sadness, and anxiety.

3.2 Matching the individual components and emotion words

The most striking finding with regard to the individual meaning components was the very high match rate between some components and related emotion words, suggesting that these components could be a crucial part of the word in question. The low match rates, in turn, may indicate several conditions, ranging from the translation or issues in the current study to the participants’ doubt that the component was in fact part of the explication.

We view our secondary data on individual meaning components from two perspectives: 1) all the meaning components of all words considered as one list, focusing on the top eight best matched components and the emotion words (Figure 4), and 2) seven sets of related individual components grouped according to the emotion words in question (Figure 5).

When considering the findings in this subsection, we must keep in mind that, unlike Section 3.1 with its “one emotion word – one explication” matches, here one single component could be part of the original explication of several emotion words. For example, the component “I’m feeling (something) bad” belonged both to *pelko* ‘fear’ and *viha* ‘anger, hatred’. As with the questions in Section 3.1, the participants were able to choose one or more answers.

Figure 4 shows the results of all the meaning components in one list, the point being that there were large differences in matching the components with the emotion words originally explicated. The Finnish versions are presented in Appendix B. The plot shows that several components had a very high match rate, six components had well above a 50% match rate, and the remaining components were connected with the original emotion word only fairly loosely or even very poorly. In the best case, more than 95% of the participants connected the component “something bad can now happen to me here” with *pelko* ‘fear’, while at the other extreme, fewer than 8% of the participants made a connection between the component “I want to do something” and *pelko* ‘fear’.⁵ Interestingly, the parts of *pelko* ‘fear’ were

⁵ The exceptionally low match rate was possibly related to the complex syntax of the component,

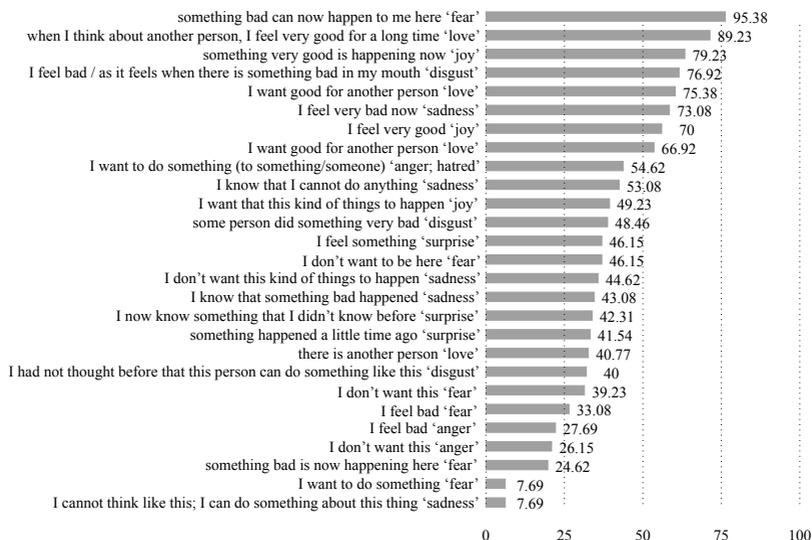


Figure 4. Findings on expected matching of individual meaning components with emotion words. The horizontal axis shows the percentage of participants ($N = 130$) making the expected match between a given component and the related word(s).

among both the best and the poorest matched individual meaning components: from 95% (124) on “something bad can now happen to me here” to 8% (10) on “I want to do something”.

Figure 5 shows seven complete sets of individual meaning components.⁶ Comparison between the sets shows how, within certain sets, there is great variation in matching the component rates (e. g. *pelko* ‘fear’ or *suru* ‘sadness’), while within some sets (e. g. *yllätys* ‘surprise’) all the components are equally

even if intuitively; thus, the match rate should have been fairly high. Based on our experience and intuitive understanding, sadness could be linked with powerlessness and incapability. In terms of the NSM, this semantic feature of sadness could be captured with phrases such as “I can’t think like this: I can do something about this thing”. We did pay attention to the potentially problematic formation of this particular component in the planning phase of the study but in the end we decided to stay with the original formulation. Reformulating and retesting this component might be worth considering in future studies.

⁶ Taken together, each of these sets would provide a suggestion for a full explication of one of the emotion words.



Figure 5. The seven sets of related individual components, grouped according to the emotion words in question. The sets are organized according to the highest percentages of matches for each emotion word. The horizontal axis shows the percentage of participants ($N = 130$) who made the expected match with a given component and the related emotion word(s).

(loosely) connected to the emotion word in question. The figure makes clearly visible how some of the components are particularly well matched with the original explication. Figure 5 allows us to compare the best meaning components of each emotion word. A complete list of the original Finnish of the translations in Figure 5 is presented in Appendix C.

4 Discussion

In the discussion, we approach our findings from three perspectives. First, we address whether or not we found empirical evidence for NSM-based meaning

descriptions of emotion vocabulary. Second, we assess our current study and make suggestions for further research. Finally, we offer some preliminary hypotheses on how NSM-based emotion explications could be used outside linguistics, especially in psychology.

4.1 Empirical evidence for NSM-based meaning components of emotion vocabulary

Our study was motivated by the forty-year-old linguistic debate over empirical evidence of meaning components of emotion words. So, did we find the evidence? The answer is both yes and no. With some (but not too many) words, explications, and individual meaning components, there were very high match rates, suggesting that the original NSM-based explications or some parts of them could be closely connected with the emotion words they have been explicating. With the majority of the explications and components, however, the match rate was relatively moderate, and in some cases it was even poor. Interestingly, our participants seemed to be quite certain of their answers: they very seldom chose the answer “I don’t know” (the percentage with *pelko* ‘fear’, *viha* ‘anger, hatred’, *rakkaus* ‘love’ and *ilo* ‘joy’ was in fact 0%; ‘surprise’ was the big exception, with 14% of participants choosing “I don’t know”).

The most provocative result of the present paper pertains to the individual meaning components. The case of ‘fear’ was especially interesting: 95% (124) of the participants found that the component “something bad can now happen to me here” is connected with *pelko* ‘fear’, while just 8% (10) found the component “I want to do something” connected with ‘fear’. These findings suggest that the component “something bad can now happen to me here” is a dominant part of the negative feeling of ‘fear’. These findings suggest that the component “something bad can now happen to me here” is a dominant part of the negative feeling of ‘fear’. In other words, this component seems to be a part of not only the English word *fear* but also the Finnish word *pelko* ‘fear’. In contrast, the component “I want to do something” does not seem to be an actual meaning component of *pelko* ‘fear’ in Finnish, while it may be a component of English *fear*.

Despite the translation and adaptation process from English to Finnish, some meaning components were very strongly connected with certain emotion words. This may also indicate the universality of some individual meaning components, even if the full explications of the emotion in question are

language-dependent. It is to be noted that our current study does not provide any empirical evidence for the meaning component in the English language. The findings from our study can only suggest support for some highly accepted components or full explications. Especially in the situation of a translated (English to Finnish) study setting, we cannot even think about falsifying any of the original English explications.

As reported in our findings, the results show a great deal of variation. First of all, there was variation between the participants' responses. Second, there was variation within the responses of the individual participants. This calls for further discussion, which is beyond the scope of this paper. The variation may be linked to various aspects, such as the distinguishing power of NSM as a method for semantic analysis (Tissari et al. 2019; see also Gladkova et al. 2016), and the completeness and distinctiveness of the original explications. Also, we need to be aware of co-existence and the mixing of emotions (e. g. Tissari 2011). A person in love is often a happy person, too.

Naturally, the findings presented in this paper could be analysed in multiple ways. For example, one could focus on the variety of unexpected answers or one could seek to find which were the most popular combinations of choices when the participants selected more than one answer.

4.2 Assessment of the current study and suggestions for further settings

The very idea of the current study, a question that has been asked over and over again during past decades, seems simple: do laypeople recognize NSM-based explications of emotion words? Combined with an empirical survey, statistical analysis, and a language other than English, the question becomes very complicated. After all the translating, adjusting, and deconstructing, what is ultimately the actual phenomenon to be tested? In our view, the overall study setting worked out well. The survey form was constructed and carefully piloted with the help of a professional statistician, and the data were collected in a controlled situation.

Regarding some details on the questionnaire form, we had to make compromises to find the best possible solution. We were well aware of the few slightly unnatural formulations, and we took these cases into account when analysing and discussing the results. Especially in the case of *viha* 'anger, hatred', we noted the linguistic and cultural differences between Finnish and English. In the case of *inho* 'disgust', a friendly colleague pointed out a minor

translation error in the form; this is regrettable, although not fatal. In addition to the seven emotion words used as the focus of this study, there were seven filler words among the answers. Most of these referred to different emotions, while some had meanings fairly similar to the words studied, which caused some synonymy in the setting.

One possible concern is related to the alphabetical order of the emotion words in the questionnaire, with *ahdistus* ‘anxiety’ being the first on the list. As seen in the findings, the filler word *ahdistus* ‘anxiety’ received many answers. Was this because the informants really meant to choose it or was it just because of its place at the beginning of the list? This question cannot be clarified in hindsight. Yet, taking this into account would possibly not have been an option, since randomizing the word list for each questionnaire page would have slowed the answer process too much. One option would be to have the word list randomized for each participant separately; unfortunately, the survey program that we used did not allow this. Having shared all this information, however, our understanding is that these issues did not ultimately have a significant effect on the study.

In future studies the current setting could be repeated with native speakers of English and an English-language questionnaire; this would enlighten the universal aspect of the emotion vocabulary.⁷ Also, it would be interesting to focus on the unexpected matches – those components which were not matched in an expected way and why. From the current perspective, and after seeing the quantitative data, obtaining open-ended answers would be very attractive. We carefully considered this option at the beginning of our study, but dropped the plan because of statistical challenges with qualitative data.

4.3 Could our findings be used in psychology or other fields outside linguistics?

There is a vast body of research literature on the intersection of linguistics and psychology, which notes that the language we use affects our experience of emotions (see, for example, Russell et al. 1995; Wierzbicka 1999; Harkins & Wierzbicka 2001; Fontaine et al. 2013; Kövecses 2005). Language is also the primary method for analysing our emotional experiences (e. g. Barrett et al. 2007: 374–377). There are still major challenges related to emotion analysis, two of which are closely related to the topics studied in the NSM

⁷ Open data available at <https://github.com/KimmoVehkalahti/Emotions> (accessed 2020-04-17).

framework. One is the circularity of meaning descriptions, which means that the term being defined is ultimately used as a part of the definition by creating a definitional chain: a word is explained by other words, which are explained by using the first word. Another is the language dependence of emotion concepts. Findings in the current study and NSM in general may help people in describing or analysing emotions from at least three perspectives.

First, semantic primes allow people to talk about emotions in depth without using the more polysemous everyday word for emotions. This may be helpful in numerous situations: for example, with people who feel that “emotion labels” do not fit the actual feeling they are experiencing, or when someone for various reasons does not recognize or master the standard language emotion vocabulary. The challenges of explaining emotion word content are not limited to cross-linguistic situations. Instead, problems often occur in interaction between speakers of the same language. If a couple in marital therapy, for example, could address dozens of possible unambiguous non-circular meaning components of love, instead of talking about love in conventional terms, would this bring their discussion to the next level?

Second, based on the results from earlier NSM studies and the empirical findings in this study, it seems very likely that recognizable key components are closely connected to each emotion concept; for example, fear-like concepts include the elements “thinking that bad things can happen” and “feeling bad because of it”, or joy-like concepts which include the elements “something very good is happening now” and “feeling very good”. We want to emphasize here that we are not supposing that these kinds of components would make a full or exhaustive explication of the concept of fear. Rather, we would say that it is hard to think that these suggested components of ‘fear’, published in previous studies and empirically confirmed in this paper, would *not* be part of the central meaning of *pelko* ‘fear’ in Finnish or English.

Third, NSM-based meaning components may be more or less universal, as different languages may conceptualize and verbalize emotions differently. The meta-language used to make the elements, however, makes it possible to compare the conceptualizations and to find similarities and differences between languages or cultures. Considering that these elements are: a) written in a non-circular and non-technical meta-language and b) are fairly well (if not fully) translatable to many (if not all) languages, this remarkable notion has gained empirical evidence, and it is of great interest today in cross-cultural psychology and psychiatry.

Finally, the overall results of this study support earlier empirical findings

(Gladkova et al. 2016; Vanhatalo & Torkki 2018) on laypeople's general ability to operate with NSM explications. Besides contributing to the theoretical assumptions on laypeople's ability to conceive NSM concepts, this notion emphasizes the practical value of NSM in helping laypeople to analyse lexical meaning in various contexts (such as earlier studies on pain; see Wierzbicka 2012). It must be kept in mind, however, that making a complex piece of language (such as an emotion or a form or instruction) simple is not necessarily easy as a process; it may require a great deal of effort by professional semanticists. An intriguing meta-methodological and philosophical question is whether we can trust the semantic assessments given by laypeople. Are not experienced and well-trained semanticists much more likely to capture the meaning components of a word? In our understanding, explications created by professionals must be tested by non-linguists. Here we agree, for example, with Ye (2013: 404), who claims that ordinary people's language use gives us access to their "raw experience" of emotion (see Goddard 2002: 19–24).

5 Conclusion

Our hypothesis is that NSM could be used to explain emotion words in both research and intervention settings, involving laypeople and cross-linguistic perspectives. An especially promising aspect is related to speakers needing easy languages. There is a great need for methods to talk about emotions among people with cognitive disabilities, memory disorders, and other language-related problems. The need for simple language with which one can discuss complex issues, including emotions, is also well recognized by those working with immigrants and speakers of second languages. The option of identifying meaning components of emotion concepts instead of trying to deal with cross-linguistic (mis)matches of emotion vocabulary might crucially improve mutual understanding across languages and cultures, for instance, in therapy sessions. We are calling for new test settings, combining psychology and linguistics, to explore these ideas further and assess these possibilities.

Appendix A

The original NSM explications chosen for this study, followed by the translated and adjusted Finnish versions used in the questionnaire.

surprised ‘someone was **surprised**’ (Goddard 2015: 298)

this someone X thought something about something at that time
at the same time this someone felt something because of it
a short time before it was like this:

– something happened

– because of this, this someone knew something about something

after this, this someone thought about it like this:

“I didn’t know before that it will be like this

I know it now”

when this someone thought like this, this someone felt something

like people feel at many times when they think like this

vähän aikaa sitten tapahtui jotain ‘something happened a little time ago’

tämän takia tiedän nyt jotain ‘because of that, I know something now’

tätä en ennen tiennyt ‘I didn’t know this before’

kun ajattelen tätä minusta tuntuu joltakin ‘when I think about this, I feel something’

fear-like emotions (Wierzbicka 2014: 78)

It can be like this:

Someone thinks like this:

“Something bad can happen to me here now

I don’t want this

I don’t want to be here

I want to do something because of this”

When this someone thinks like this, this someone feels something bad because of this, like people often feel when they think like this

minulle voi tapahtua täällä nyt jotain pahaa ‘something bad can now happen to me here’

en halua tätä ‘I don’t want this’

en halua olla täällä ‘I don’t want to be here’

tämän takia haluan tehdä jotain ‘I want to do something because of this’

kun ajattelen tätä, minusta tuntuu pahalta ‘when I think about this, I feel bad’

anger-like and wut-like emotions (Wierzbicka 2014: 83)

It can be like this

Someone thinks like this:

“Something bad is happening here now

I don’t want this

I want to do something (to something) because of this”

When this someone thinks like this, this someone feels something bad because of this.

täällä tapahtuu nyt jotain pahaa ‘something bad is now happening here’

en halua tätä ‘I don’t want this’

haluan tehdä jotakin (jollekin) tämän takia ‘I want to do something (to something/someone) because of this’

minusta tuntuu pahalta tämän takia ‘I feel bad because of this’

joy (X felt joy) (Wierzbicka 1999: 50)

X felt something because X thought something

Sometimes a person thinks:

“Something very good is happening

I want this to be happening”

When this person thinks this, this person feels something very good

X felt something like this

Because X thought something like this

nyt tapahtuu jotain oikein hyvää ‘something very good is happening now’

haluan, että tällaista tapahtuu ‘I want this kind of things to happen’

minusta tuntuu oikein hyvältä, kun ajattelen tätä ‘when I think about this I feel very good’

sad / sadness (Goddard 2011: 112)

Someone X is sad at this time:

a. someone X thinks like this at this time:

“I know that something bad happened

I don’t want things like this to happen

I can’t think like this: I can do something because of this

I know that I can’t do anything”

b. because of this, this someone feels something bad at this time

c. like people often feel when people think like this

tiedän, että jotain pahaa tapahtui ‘I know that something bad happened’

en halua, että tällaisia asioita tapahtuu ‘I don’t want this kind of thing to happen’

en voi ajatella näin: tälle asialle voin tehdä jotain ‘I can’t think like this: I can do something about this thing’

tiedän, että en voi tehdä mitään ‘I know that I can’t do anything’

tämän takia minusta tuntuu nyt pahalta ‘I feel bad because of this’

disgust (Goddard 2014: 81)

Someone X is disgusted (with someone) at this time.

a. someone X thinks like this about someone at this time:

“this someone did something very bad

I didn't think before that this someone can do something like this”

b. because of this, this someone feels something very bad at this time

c. like someone can feel when something bad happens in this someone's mouth [m]

because there is something very bad inside this someone's mouth [m]

d. this someone can't not feel like this

eräs ihminen teki jotain oikein pahaan ‘some person did something very bad’

en ollut ennen ajatellut, että tämä ihminen voi tehdä jotain tällaista ‘I had not thought before that this person could do something like this’

tämän takia minusta tuntuu pahalta ‘I feel bad because of this’

niin kuin minusta tuntuu silloin kun suussani on jotain pahaan ‘like I feel when there is something bad in my mouth’

rakkaus ‘love’ (Tuovila 2005: 112, translated into English by the authors)

X tuntee jotakin, koska hän ajattelee jotakin ‘X feels something, because s/he thinks about something’

joskus ihminen ajattelee ‘sometimes a person thinks’

ajattelen: on olemassa toinen ihminen ‘I think: there is another person’

kun ajattelen häntä, minusta tuntuu erittäin hyvältä ‘when I think about her/him, I feel very good’

haluan hänelle hyvää ‘I want good for her/him’

kun ihminen ajattelee tätä, hänestä tuntuu hyvin hyvältä kauan aikaa ‘when a person thinks this, s/he feels good for a long time’

X tuntee jotakin tämän kaltaista, koska hän ajattelee jotakin tämän kaltaista ‘X feels something like this, because s/he thinks something like this’

on olemassa toinen ihminen ‘there is another person’

kun ajattelen häntä, minusta tuntuu erittäin hyvältä ‘when I think about her/him, I feel very good’

haluan hänelle hyvää ‘I want good for her/him’

kun ajattelen tätä, minusta tuntuu hyvin hyvältä kauan aikaa ‘when I think about this, I feel good for a long time’

Appendix B

The match rate of individual meaning components and originally explicated emotion words. The percentage refers to the number of participants making the expected match. The list also presents all the individual components used in this study. The Finnish version is provided in brackets.

95.38 something bad can now happen to me here (*minulle voi tapahtua täällä nyt jotain pahaan*), ‘fear’

89.23 when I think about another person, I feel very good for a long time (*kun ajattelen toista ihmistä, minusta tuntuu erittäin hyvältä kauan aikaa*), ‘love’

- 79.23 something very good is happening now (*nyt tapahtuu jotain oikein hyvää*), ‘joy’
- 76.92 I feel bad / as it feels when there is something bad in my mouth (*minusta tuntuu pahalta / niin kuin tuntuu silloin kun suussani on jotain paha*), ‘disgust’
- 75.38 I want good for another person (*haluan hyvää toiselle ihmiselle*), ‘love’
- 73.08 I feel very bad now (*minusta tuntuu nyt pahalta*), ‘sadness’
- 70 I feel very good (*minusta tuntuu oikein hyvältä*), ‘joy’
- 66.92 I want good for another person (*haluan toiselle ihmiselle hyvää*), ‘love’
- 54.62 I want to do something (to something/someone) (*haluan tehdä jotakin (jollekin)*), ‘anger, hatred’
- 53.08 I know that I cannot do anything (*tiedän, että en voi tehdä mitään*), ‘sadness’
- 49.23 I want this kind of things to happen (*haluan, että tällaista tapahtuu*), ‘joy’
- 48.46 some person did something very bad (*eräs ihminen teki jotain oikein paha*), ‘disgust’
- 46.15 I feel something (*minusta tuntuu joltakin*), ‘surprise’
- 46.15 I don’t want to be here (*en halua olla täällä*), ‘fear’
- 44.62 I don’t want this kind of things to happen (*en halua, että tällaisia asioita tapahtuu*), ‘sadness’
- 43.08 I know that something bad happened (*tiedän, että jotain paha tapahtui*), ‘sadness’
- 42.31 I now know something that I didn’t know before (*tiedän nyt jotain, jota en ennen tiennyt*), ‘surprise’
- 41.54 something happened a little time ago (*vähän aikaa sitten tapahtui jotain*), ‘surprise’
- 40.77 there is another person (*on olemassa toinen ihminen*), ‘love’
- 40 I had not thought before that this person can do something like this (*en ollut ennen ajatellut, että tämä ihminen voi tehdä jotain tällaista*), ‘disgust’
- 39.23 I don’t want this (*en halua tätä*), ‘fear’
- 33.08 I feel bad (*minusta tuntuu pahalta*), ‘fear’
- 27.69 I feel bad (*minusta tuntuu pahalta*), ‘anger’
- 26.15 I don’t want this (*en halua tätä*), ‘anger’
- 24.62 something bad is happening now here (*täällä tapahtuu nyt jotain paha*), ‘anger’
- 7.69 I want to do something (*haluan tehdä jotain*), ‘fear’
- 7.69 I cannot think like this: I can do something about this thing (*en voi ajatella näin: tälle asialle voin tehdä jotain*), ‘sadness’

Appendix C

The match rates of sets of individual meaning components and the originally explicated emotion words, grouped according to the emotion words. The percentage refers to the number of participants making the expected match.

‘fear’ (pelko)

- 95.38 something bad can now happen to me here (*minulle voi tapahtua täällä nyt jotain paha*)
- 46.15 I don’t want to be here (*en halua olla täällä*)

- 39.23 I don't want this (*en halua tätä*)
 33.08 I feel bad (*minusta tuntuu pahalta*)
 7.69 I want to do something (*haluan tehdä jotain*)

'love' (*rakkaus*)

- 89.23 when I think about another person, I feel very good for a long time
 (*kun ajattelen toista ihmistä, minusta tuntuu erittäin hyvältä kauan aikaa*)
 75.38 I want good for another person (*haluan hyvää toiselle ihmiselle*)
 66.92 I want good for another person (*haluan toiselle ihmiselle hyvää*)
 40.77 there is another person (*on olemassa toinen ihminen*)

'joy' (*ilo*)

- 79.23 something good is happening now (*nyt tapahtuu jotain oikein hyvää*)
 70 I feel very good (*minusta tuntuu oikein hyvältä*)
 49.23 I want this kind of things to happen (*haluan, että tällaista tapahtuu*)

'disgust' (*inho*)

- 76.92 I feel bad / as it feels when there is something bad in my mouth (*minusta tuntuu pahalta / niin kuin tuntuu silloin kun suussani on jotain pahaa*)
 48.46 some person did something very bad (*eräs ihminen teki jotain oikein pahaa*)
 40 I had not thought before that this person can do something like this
 (*en ollut ennen ajatellut, että tämä ihminen voi tehdä jotain tällaista*)

'sadness' (*suru*)

- 73.08 I feel bad now (*minusta tuntuu nyt pahalta*)
 53.08 I know that I cannot do anything (*tiedän, että en voi tehdä mitään*)
 44.62 I don't want these kinds of things to happen (*en halua, että tällaisia asioita tapahtuu*)
 43.08 I know that something bad happened (*tiedän, että jotain pahaa tapahtui*)
 7.69 I cannot think like this: I can do something about this thing (*en voi ajatella näin: tälle asialle voin tehdä jotain*)

'anger; hatred' (*viha*)

- 54.62 I want to do something (to someone/something) (*haluan tehdä jotakin (jollekin)*)
 27.69 I feel bad (*minusta tuntuu pahalta*)
 26.15 I don't want this (*en halua tätä*)
 24.62 something bad is happening now here (*täällä tapahtuu nyt jotain pahaa*)

'surprise' (*hämmästyks*)

- 46.15 I feel something (*minusta tuntuu joltakin*)
 42.31 I now know something that I did not know before (*tiedän nyt nyt jotain, jota en ennen tiennyt*)
 41.54 something happened a little time ago (*vähän aikaa sitten tapahtui jotain*)

References

- Barrett, Lisa & Mesquita, Batja & Ochsner, Kevin & Gross, James. 2007. The experience of emotion. *Annual Review of Psychology* 58. 373–403.
- Barrett, Lisa & Russell, James A. 1998. Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology* 74. 967–984.
- Ekman, Paul. 1993. Facial expression and emotion. *American Psychologist* 48(4). 384–392.
- 1999a. Basic emotions. In Dalgleish, T. & Power, M. J. (eds.), *Handbook of cognition and emotion*, 45–60. New York: John Wiley & Sons.
- 1999b. Facial expressions. In Dalgleish, T. & Power, M. J. (eds.), *Handbook of cognition and emotion*, 301–320. New York: John Wiley & Sons.
- 2016. What scientists who study emotion agree about. *Perspectives on Psychological Science* 11(1). 31–34.
- The Finnish Association for Mental Health. 2019. *Tunteet* [Emotions]. (<https://mieli.fi/fi/mielenterveys/itsetuntemus/tunteet>) (Accessed 2019-09-15).
- Fontaine, Johnny & Scherer, Klaus & Soriano, Cristina. 2013. *Components of emotional meaning: A sourcebook*. Oxford: Oxford University Press.
- Gladkova, Anna & Vanhatalo, Ulla & Goddard, Cliff. 2016. The semantics of interjections: An experimental study with Natural Semantic Metalanguage. *Applied Psycholinguistics* 37(4). 841–865.
- Goddard, Cliff. 2002. Explicating emotions across languages and cultures: A semantic approach. In Fussell, S. R. (ed.), *The verbal communication of emotions: Interdisciplinary perspectives*, 19–53. Mahwah, NJ: Lawrence Erlbaum.
- 2008. Natural Semantic Metalanguage: The state of the art. In Goddard, Cliff (ed.), *Cross-linguistic semantics*, 1–34. Amsterdam: John Benjamins.
- 2011. *Semantic analysis: A practical introduction*. 2nd edn. Oxford: Oxford University Press.
- 2014. On “disgust”. In Baider, Fabienne & Cislaru, Georgeta (eds.), *Linguistic approaches to emotions in context*, 73–98. Amsterdam: John Benjamins.
- 2015. The complex, language-specific semantics of “surprise”. *Review of Cognitive Linguistics* 13(2). 291–313. DOI: 10.1075/rcl.13.2.02god.
- Goddard, Cliff (ed.). 2018. *Minimal English for a global world: Improved communication using fewer words*. Cham: Palgrave Macmillan.
- Goddard, Cliff & Wierzbicka, Anna. 2014. *Words and meanings: Lexical semantics across domains, languages and cultures*. Oxford: Oxford University Press.
- Harkins, Jean. 2001. Talking about anger in Central Australia. In Harkins, Jean & Wierzbicka, Anna (eds.), *Emotions in crosslinguistic perspective*, 197–215. Berlin: Mouton de Gruyter.
- Harkins, Jean & Wierzbicka, Anna (eds.). 2001. *Emotions in crosslinguistic perspective*. Berlin: Mouton de Gruyter.

- Jack, Rachael & Garrod, Oliver & Schyns, Philippe. 2014. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology* 24(2). 187–192. DOI: 10.1016/j.cub.2013.11.064.
- Kövecses, Zoltán. 2005. *Metaphor in culture: Universality and variation*. Cambridge: Cambridge University Press.
- Ortony, Andrew & Turner, Terence. 1990. What's basic about basic emotions? *Psychological Review* 97(3). 315–331.
- Russell, James A. & Fernández-Dols, José-Miguel & Manstead, Antony S. R. & Wellenkamp, J. C. (eds.). 1995. *Everyday conceptions of emotion: An introduction to the psychology, anthropology and linguistics of emotion*. Dordrecht: Kluwer.
- Scarantino, Andrea. 2014. Basic emotions, psychological construction, and the problem of variability. In Barrett, Lisa & Russell, James A. (eds.), *The psychological construction of emotion*, 334–376. Amsterdam: Guilford Publications.
- Tissari, Heli. 2011. Oops blush! Beyond metaphors of emotion. In Gries, S. T. & Brdar, M. & Zic Fuchs, M. (eds.), *Convergence and expansion in cognitive linguistics*, 291–304. Amsterdam: John Benjamins.
- Tissari, Heli & Vanhatalo, Ulla & Siirainen, Mari. 2019. From corpus-assisted to corpus-driven NSM explications: The case of Finnish *viha* (anger, hate). *Lege Artis: Language Yesterday, Today, Tomorrow* 4. 290–334.
- Tuovila, Seija. 2005. *Kun on tunteet: Suomen kielen tunnesanojen semantiikkaa* [Because we have emotions: On the semantics of Finnish words for emotions]. Oulu: University of Oulu, Department of Finnish, Information Studies and Logopedics.
- Vanhatalo, Ulla & Torkki, Juhana. 2018. Introducing the concept of the ‘65 words’ to the public in Finland. In Goddard, Cliff (ed.), *Minimal English for a global world: Improved communication using fewer words*, 225–258. Cham: Palgrave Macmillan.
- Watson, David & Tellegen, Auke. 1985. Toward a consensual structure of mood. *Psychological Bulletin* 98. 219–235.
- Wierzbicka, Anna. 1992a. Defining emotion concepts. *Cognitive Science* 16. 539–581.
- 1992b. *Semantics, culture and cognition: Universal human concepts in culture-specific configurations*. Oxford: Oxford University Press.
- 1998. “Sadness” and “anger” in Russian: The non-universality of the so-called “basic human emotions”. In Athanasiadou, Angeliki & Tabakowska, Elzbieta (eds.), *Speaking of emotions: Conceptualisation and expression*, 3–28. Berlin: Mouton de Gruyter.
- 1999. *Emotions across languages and cultures: Diversity and universals*. Cambridge: Cambridge University Press.
- 2012. Is pain a human universal? A cross-linguistic and cross-cultural perspective on pain. *Emotion Review* 4(3). 307–317. DOI: 10.1177/1754073912439761.

- 2014. *Imprisoned in English: The hazards of English as a default language*. Oxford: Oxford University Press.
- Wood, Alex & Johnson, Judith (eds.). 2016. *The Wiley handbook of positive clinical psychology: An integrative approach to studying and improving well-being*. Hoboken, NJ: Wiley-Blackwell.
- Wood, Alex & Tarrier, Nicholas. 2010. Positive clinical psychology: A new vision and strategy for integrated research and practice. *Clinical Psychology Review* 30. 819–829.
- Ye, Zhengdao. 2013. Comparing the Natural Semantic Metalanguage (NSM) approach to emotion and the GRID paradigm. In Fontaine, Johnny & Scherer, Klaus & Soriano, Cristina (eds.), *Components of emotional meaning: A sourcebook*, 399–409. Oxford: Oxford University Press.

Contact information:

Ulla Vanhatalo
Department of Languages
P. O. Box 24 (Unioninkatu 40)
00014 University of Helsinki
Finland
e-mail: ulla(dot)vanhatalo(at)helsinki(dot)fi

Heli Tissari
e-mail: heli(dot)tissari(at)helsinki(dot)fi

Taru Lilja
HAMK Edu Research Unit
Häme University of Applied Sciences
P. O. Box 230
13101 Hämeenlinna
Finland
e-mail: taru(dot)lilja(at)hamk(dot)fi

Kimmo Vehkalahti
Centre for Social Data Science
P. O. Box 18 (Unioninkatu 35)
00014 University of Helsinki
Finland
e-mail: kimmo(dot)vehkalahti(at)helsinki(dot)fi

Mari Siirainen
Department of Digital Humanities
P. O. Box 3 (Fabianinkatu 33)
00014 University of Helsinki
Finland
e-mail: mari(dot)siirainen(at)helsinki(dot)fi

Fassi Fehri, Abdelkader. 2018. *Constructing feminine to mean: Gender, number, numeral, and quantifier extensions in Arabic*. London: Rowman & Littlefield. Pp. 233.

Reviewed by Abdel Rahman Mitib Altakhaineh and Aseel Zibin

1 Introduction

The grammatical category ‘feminine’ has been viewed in various ways in the relevant literature, typically as the marked member of the category Gender (henceforth Gen), i. e. denoting sex (Kibort & Greville 2008) and/or animacy (Dahl 2000). It enters into an Agree relation in Chomsky’s work (1995; 2000) and as such is interpretable on the controller or the original locus but uninterpretable on its target (as in e. g. Moravcsik 1988). In this regard, interpretable or valued features make a semantic contribution to the interpretation of an item at an interface level, while uninterpretable or unvalued ones do not. Yet, the studies that have tackled Gender have not clearly explained why it is that not only general categories, e. g. nouns, verbs, adjectives and adverbs, but also numerals, pluralities, singularities and quantifiers are ‘feminisable’. Fassi Fehri’s monograph is the first attempt to provide an answer to the above question. In addition, the author furnishes a systematic account of the feminine in Arabic, as well as its use in Afro-Asiatic, Germanic, Slavic, and Romance languages.

The author begins by providing an overview of the main themes discussed in his book. The feminization of numerals, singulatives,¹ pluratives,² and quantifiers is examined first. Next, an account of the constructional nature of Gender is set out, explaining the way in which this latter is not inherent, but must actively be built into the nominal domain (see Alexiadou 2004; Kihm 2005; Lowenstamm 2008). Thirdly, he discusses *Unity* which is, according to him, a process that creates a unit either by packaging things or by taking several individuals or objects and putting them together to create a new unit. He also sheds more light on the role of unity in the grammar of individuation and Number (henceforth Num; Fassi Fehri 2003). The structure of Quantifier

¹ The *singulative* is a process through which a collective is changed into a single unit or individual, commonly marked via Gender (-at) triggering feminine singular agreement on its target (p. 7).

² The *plurative* is a process through which a collective noun phrase (NP) is changed into a group unit or a collection unit resulting in an integrated whole. It is morphologically marked on the controller, the target or both using the same feminine suffix as the singulative (p. 10).

expressions and the peculiar practice of counting in the feminine in Arabic are then discussed in detail. Finally, Fassi Fehri illustrates the different projections and labels of Gen, its multiple uses and senses, as well as its role as a Classifier, and how it plays a part in the taxonomy of numbers in Arabic and Hebrew.

2 Summary of the chapters

The book comprises five chapters. In Chapter 1, based on empirical data from both Standard Arabic (SA) and Moroccan Arabic (MA), Fassi Fehri demonstrates that Gender is more active not only in the Noun Phrase (NP) or Determiner Phrase (DP) structure but also in the upper Complementizer Phrase (CP) structure. This suggests that Gender is found in various layers, which are projections of head categories, e. g. verbs, nouns, prepositions, adjectives, etc., and not uniquely in nouns (N). The analysis illuminates the semantic diversity of Gender over and above the narrower scope of sex differentiation to include *individuation* (e. g. singulative vs. plurative), *collectivity*, *quantity*, *abstractness*, *size*, *evaluation* and *perspectivization*. Relying on the structure of the layer as well as on whether it is interpretable or non-interpretable, Gen involves various features and values, including [+indiv], [+fem], [+group], [+endearing], [+small/big], or [+good/bad].

In Chapter 2, Fassi Fehri explains that Gender in Arabic co-occurs with Number, suggesting that typologies such as *classifier languages*, *gender languages*, or *number languages* are no longer feasible. This is because Gender may co-occur with Number in other languages such as the Romance languages and in Hebrew, and Berber, which would necessitate a new typology. Instead, Fassi Fehri adopts a multi-layered and polysemous view of Gender, proposing that Gen (and typically the feminine) cannot be restricted only to the noun (N) domain as has generally been posited in dominant analyses of Indo-European gender. It is in fact *hyperonymic*, a general category that integrates more diverse and structurally organized meanings found cross-linguistically, with sex/animacy only a *hyponymic* or a special case. Fassi Fehri also argues that this multi-layered analysis can be extended to languages other than Arabic, including Hebrew, Berber, and the Romance languages.

In Chapter 3, Fassi Fehri first argues against a simple derivation of numerals based on Merge (cf. Chomsky 2008). Dispensing with the Merge analysis explains why 3 ('three') in Semitic is 3-GROUP or 3-SET which is

directly countable and collective, suggesting that 3 exhibits some individual or atomic properties making it indirectly countable. On the other hand, 3 in Slavic languages has some additional peculiarities because it uses Gen (neutral, fem and variable) to distinguish three kinds of collective numerals. These are identified as *n*-numerals (for counting numbers), *c*-numerals (for counting objects), and *o*-numerals (for ranking objects). Adopting a *root-category* model (Marantz 2005; Borer 2005; Harley 2014), Fassi Fehri proposes then that numerals are “born” as linguistic expressions of *number* of type N, i. e. are created for counting or *cardinalizing*. As such, the numerosity sense of numerals is at the Root, whereas their other sense is compositionally derived via categorization, i. e. noun (N), adjective (A), verb (V), and preposition (P), on the one hand, and other combinations, e. g. Num and Gen, on the other. One of the main contributions of this third chapter is its demonstration of how *Gen polarity* is characteristic of Semitic languages, but does not occur in Germanic, Slavic, or Romance languages. Fassi Fehri argues that this polarity is better accounted for as a rule of pronunciation rather than as a switch gender rule.

Chapter 4 focuses on the means by which the inflectional ingredients within quantifier extensions, e. g. Gender, Number, Definiteness, etc. work together to construct the various interpretations of Arabic quantifiers. This means that Gender appears as $[\pm\text{fem}]$ or $[\pm\text{unit}]$ in Quantifier Phrases. The $[\pm\text{unit}]$ Gen matching found in the Quantifier Phrase (QP) is subject to the Gen polarity constraint, while the $[\pm\text{fem}]$ is governed by Probe-Goal Agree. This specifically demonstrates how the QP as well as other functional elements in the DP architecture are built and compositionally interpreted. Taking into account that a single vocabulary counterpart in Arabic, i. e. *kull* expresses English universal quantification (*all*, *every*, and *each*), Fassi Fehri identifies a trilogy of Distributive Quantifier patterns and meanings linked to the quantifier *kull*: the *kull* (all) type, the *kull* (each) type, and the *kull* (every) type. While all types are conceivably analyzable as forms of Partitive Phrase (PartP) structures, they do have differences. The *kull_{al}* type is a PartP in which the whole and the part are definite. *kull_{ea}* is a PartP in which the second member is definite, whereas the first is indefinite. In *kull_{ev}*, both its Q and its complement are indefinite. As such, if the latter is analyzed as PartP, it has a pseudo-partitive structure rather than a true partitive structure.

While the author provides a thorough analysis of the three types of quantifiers that can contribute to our understanding of the internal syntax and partly semantics of quantifiers cross-linguistically, he pays less attention to other quantifiers such as *ʔağlab* ‘most’, *ḥaʕd* ‘some’, *jamiif* ‘all’, and *ʔaktar*

‘more’. Additional analysis of the scope and semantics of the quantifiers discussed (and of others omitted) would be helpful, as well discussion of how this can cause ambiguous readings.

The final chapter offers a description of the most prominent properties of Arabic numbers based on a new theory of Number. The unique contribution of this chapter is that it introduces the notion of plurative. While the majority of theories around Number have focused on its grammatical facets, including singular, plural and dual (designated by the author as the *atomic* function of Number), this chapter is centered on the less-investigated aspects of singulatives and pluratives (see Fassi Fehri and Vinet 2008) by integrating what he calls the unity property. It is proposed that some singularities and pluralities are atomicities (e. g. *kalb* ‘dog’, *rijaal* ‘men’), while others are unities (e. g. *tuffaah-at* ‘apple-unit’, *najjaar-at* ‘carpenters as a group’). Unities are grammatically marked as feminine, and realized as singulatives for singulars and pluratives for plurals. Atomicities and unities project as AtomP and UnitP, respectively (splitting Borer’s DivP), subsuming two crucial senses of traditional classifier phrases. Employing [+atom] and [+unit] features, the author then establishes four number classes: (1) singulative = [+atom; +unit]; (2) singular = [+atom; -unit]; (3) plural = [-atom; -unit]; and (4) plurative = [-atom; +unit]. Such an elaboration of the grammatical notion of individuation may account for the many ways of numbering and counting. Feminine then differentiates between two classes of counted entities and numerals: (a) natural members and (b) objects, thus providing an answer to one of the most challenging questions in dealing with Gender.

3 Conclusion

Despite the fact that Fassi Fehri includes significant analyses from different language groups, including Romance, Slavic, Germanic, etc. in relation to Gender and individuation, more data from these languages is needed to confirm the proposed analyses. Additionally, even though the greater part of the empirical analysis is built on an examination of Standard Arabic and Moroccan Arabic, the author argues that his analysis can be extended to cover a number of diverse languages. Fassi Fehri’s investigation may not be applicable, however, to some varieties of Arabic. In Jordanian Arabic (JA), for instance, unlike in Moroccan Arabic, the plurative *najjaar-ah* ‘carpenters’ is not used, instead of which the sound plural *najjaareen* ‘carpenters’ (p. 149)

is used. The former denotes the plural form for ‘female carpenter’ in JA. In addition, JA uses neither endearing Gen *-at* nor the feminine *-ii* for the imperative verb (pp.28–30). For these reasons, in spite of its title, this monograph does not fully achieve the claimed comprehensive coverage of femininity in Arabic.

The greatest value of this monograph lies in its thorough investigation of Gender, and how it is indeed interpretable on all nominal categories and other categories in Arabic and possibly in other Afro-Asiatic, Romance, Germanic, and Slavic languages. Its importance also stems from the author’s attempt to explicate the nature of categories such as Numeral, Number, and Quantifier, their roles, and their projections in the nominal spine and causal architecture, using less technical terms. Readers are presented with examples illustrating each argument and analysis and are guided through various approaches to interpret data which provide an explanation for the analysis adopted in this monograph. Such an approach makes this book reader-friendly for an international linguistic readership.

References

- Alexiadou, Artemis. 2004. Inflection class, gender and DP internal structure. In Müller, Gereon & Gunkel, Lutz & Zifonun, Gisela (eds.), *Exploration in nominal inflection*, 21–50. Berlin: Mouton de Gruyter.
- Borer, Hagit. 2005. *Structuring sense, volume I: In name only*. Oxford: Oxford University Press.
- Chomsky, Noam. 1995. *The minimalist program*. Cambridge, MA: MIT Press.
- 2000. Minimalist inquiries: The Framework. In Martin, R. & Michaels, D. & Uriagereka, J. (eds.), *Step by step: Essays on minimalist syntax in honor of Howard Lasnik*, 89–156. Cambridge, MA: MIT Press.
- 2008. On phases. In Freidin, R. & Otero, C. & Zubizarreta, M. L. (eds.), *Foundation issues in linguistic theory: Essays in honor of Jean-Roger Vergnaud*, 133–66. Cambridge, MA: MIT Press.
- Dahl, Östen. 2000. Elementary gender distinctions. In Unterbeck, B. & Rissanen, M. & Nevalainen, T. & Saari, M. (eds.), *Gender in grammar and cognition*, 577–593. Berlin: Mouton de Gruyter.
- Fassi Fehri, Abdelkader. 2003. Nominal classes and parameters across interfaces and levels, with a particular reference to Arabic. *Linguistic Research* 8(2). 41–108.
- Harley, Heidi. 2014. On the identity of roots. *Theoretical Linguistics* 40(3–4). 225–276.

- Kibort, Anna & Greville, Corbett. 2008. *Gender: Grammatical Features*.
(<http://www.grammaticalfeatures.net/features/gender.html>) (Accessed 2019-03-15).
- Kihm, Alain. 2005. Noun class, gender and the lexicon-syntax-morphology interfaces: A comparative study of Niger-Congo and Romance languages. In Cinque, G. & Kayne, R. S. (eds.), *The Oxford handbook of comparative syntax*, 459–512. Oxford: Oxford University Press.
- Lowenstamm, Jean. 2008. On little n, √, and types of nouns. In Hartmann, Jutta & Hegedűs, Veronika & van Riemsdijk, Henk (eds.), *Sounds of silence: Empty elements in syntax and phonology*, 105–144. Amsterdam: Elsevier.
- Marantz, Alec. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review* 22(2–4). 429–445.
- Moravcsik, Edith. 1988. Agreement and markedness. In Barlow, M. & Ferguson, C. (eds.), *Agreement in natural language: Approaches, theories, descriptions*, 89–106. Stanford: CSLI.

Contact information:

Abdel Rahman Mitib Altakhaineh
Al Ain University
P.O. Box 64141
Al Ain, UAE
e-mail: abdelrahman(dot)takhaine(at)aau(dot)ac(dot)ae

Aseel Zibin
Queen Rania St.
University of Jordan
School of Foreign Languages
P.O. Box 11942
Amman, Jordan
e-mail: a(dot)zabin(at)ju(dot)edu(dot)jo

Aikhenvald, Alexandra Y. & Dixon, R. M. W. (eds.). 2017. *Commands: A cross-linguistic typology: Explorations in linguistic typology*. Oxford: Oxford University Press. Pp. xxiv + 328.

Reviewed by Rea Peltola

1 Introduction

Alexandra Y. Aikhenvald's & R. M. W. Dixon's book *Commands: A Cross-Linguistic Typology* explores the diversity in world languages in the marking and the meaning of commands, as well as the semantic structures and cultural values and ideas underlying this variation. The array of investigated languages is illustrative of different language families, geographical areas and typological profiles, including Aguaruna and Ashaninka Satipo (both from Peru), Dyirbal (Australia), Japanese, Korowai (West Papua), Karawari (Papua New Guinea), Lao, Northern Paiute (a native language of the United States), Nungon (Papua New Guinea), Quechua, Tayatuk (Papua New Guinea), Wolaitta (Ethiopia) and Zenzontepec Chatino (Mexico). The studies draw on rich empirical data which is mostly achieved through field work.

The first extensive and systematic investigation concentrating on the typological description of directive sentences was Xrakovskij (2001). It presented results from a sample of 23 languages which were obtained through a questionnaire designed to capture comparable morphological, syntactic and semantic-pragmatic information from languages of different types. In the first chapter of the book, Birjulin & Xrakovskij (2001) provided an unconventionally broad definition of imperatives, expanding their interest, e. g., in non-second person imperatives, verbless commands, prohibitives and directives oriented toward an already on-going event. Cross-linguistic research has thereafter deepened our understanding on some of these individual aspects of imperatives (see e. g. van der Auwera et al. 2003, on the person distinctions and the nature of imperative-hortative speech act; Miestamo & van der Auwera 2007 and van der Auwera 2010, on prohibitives). Aikhenvald (2010) was the first to explore imperatives and other command strategies in a large amount of data (a set of approximately 700 grammars of different language families), instead of a sample of a more limited number of languages, with a broad scope of analysis (although the definition of imperatives was narrower than that given by Birjulin & Xrakovskij 2001). This monograph gave a thorough account of the morphological,

syntactic, semantic and pragmatic properties of directive expressions and set an empirical foundation for future studies. An up-to-date state-of-the-art concerning studies in imperative and command strategies can be found in the introduction of Van Olmen & Heinold (2017), which is itself another recent contribution in the field.

The analytical framework of the present volume is outlined in the introductory chapter by Alexandra Y. Aikhenvald, which builds upon grammatical data from 600 languages and the results presented in Aikhenvald (2010). This chapter evokes some cross-linguistic patterns of formation and usage in imperatives and commands, all the while underlining the non-universality of many of the features generally associated with imperatives. The chapters that follow demonstrate this complexity. They generally start by presenting the conventionalized, dedicated forms of command in the studied language, taking into account the recurrent formal differences between addressee-oriented (“canonical”) and other-person-oriented (“non-canonical”) imperatives. Some observations are also given regarding the non-directive uses of imperatives, for example in greetings and curses (Japanese, Chapter 8 by Nerida Jarkey) or in asking for permission and discourse-marking (Nungon, Chapter 11 by Hannah S. Sarvasy). Attention is then drawn toward other command strategies, in other words non-imperative forms (e.g. interrogative, declarative and de-subordinated clauses) used for modulating the directive force.

In the course of the book, two major domains of interest are brought to the fore: the interplay between the distinctiveness of the imperative category and its historical, formal and semantic relatedness with neighboring verb categories, on the one hand, and the culturally and socially rooted nature of commands, on the other.

2 The imperative among other grammatical categories: peculiarity and relatedness

Not all languages have a dedicated imperative form, nor do all communities make use of a clearly defined speech act of commands. In his article (Chapter 9), N. J. Enfield shows that in Lao, an isolating-analytic language that employs sentence-final particles for distinguishing clause types, it is not obvious to consider imperatives as a sentence type of its own. Imperative sentences are structurally similar to other sentence types. They may contain one of a series

of specialized sentence-final particles, in which case the semantic category of the particle determines the type of command performed (e. g. the particle *vaj2* indicates that the speaker is asking the addressee to hurry), but they may also be formed without any explicit marking.

When dealing with linguistic phenomena as cross-linguistically diverse as sentence types and other means for expressing speech acts, there is a great risk of imposing conventional analytic frames on forms that do not fit, namely analyzing as imperatives something that is not an imperative. R. M. W. Dixon's paper (Chapter 6) revisits Dyirbal verbal inflection to demonstrate that the forms labelled "positive imperative" and "negative imperative" in a previous study are better characterized as potentiality and caution verb forms. According to Dixon, in traditional Dyirbal society, acts of ordering and forbidding were not part of the life of its members, and there was therefore no well-defined speech act category of commands. The article affords a strong example of the way in which long-term, in-depth research can shed light on the originality and the richness of the resources of a given language.

When there is a dedicated imperative form in the verbal system of a language, it tends to stand out as exceptional among verb forms, in terms of the semantic distinctions it encodes and the grammatical relations it entertains with other phrasal constituents. First, imperatives may differ from other forms in number and person marking and meaning. There may be gaps in imperative paradigms: the absence of the first-person singular imperative is not uncommon. This is the case in Quechua, investigated by Willem H. Adelaar (Chapter 2), which instead displays a first-person inclusive form ('you and me'). On the other hand, in Korowai, there are more person distinctions in the imperative than in other verb paradigms, as noted by Lourens de Vries (Chapter 12). Korowai imperatives distinguish between first, second and third person forms in singular and plural, while all other verb forms only oppose the speaker (first person) with the non-speaker (second/third person) in singular and plural. This is interesting in view of the discussion concerning the non-canonicity of third-person imperatives.

Second, imperative clauses may encode syntactic relations in an unusual way. Tim Thornes presents examples from Numic languages where the subject either is non-overt or appears in an unusual form, namely in the third-person reflexive pronominal form (Chapter 7). In this language branch, transitive imperative sentences may also exhibit atypically case-marked objects.

Furthermore, imperatives are likely to deviate from other verb forms in expression of tense, aspect, modality (TAM) and evidentiality. Typically,

fewer verbal categories are marked in imperative clauses than in declarative or interrogative clauses. In Quechua, the imperative mood is incompatible with sentential affixes indicating validation and evidentiality, whereas competing verbal categories, the future tense and the potential mood, combine with these markers (Adelaar, Chapter 2). Valérie Guérin reports that, in Tayatuk, canonical imperatives formed with a bare verb stem accept none of the TAM markers found in declarative mood (at least five tenses, one aspect and three modalities) (Chapter 10). In Ashaninka Satipo, on the other hand, imperative constructions can host a number of aspectual markers but their function is not the same as in declarative constructions. Elena Mihas provides several examples of the ways in which aspectual and modal markers are used in Ashaninka Satipo imperatives to convey pragmatic information, namely to adjust the force of a command (Chapter 4). For example, when associated with the imperative, the semelfactive marker *apaint* ‘once’ gives rise to a diminutive reading: the addressee is requested to perform the action expressed by the verb to a lesser extent than normally expected.

In all its exceptionality, imperative is nevertheless formally, semantically and historically related to certain other verb categories, in a cross-linguistically consistent way. Simon E. Overall observes a formal, most likely motivated overlap between imperative, vocative and interrogative in Aguaruna and suggests that there may be a wider category of “addressee-oriented forms” (Chapter 3). The close connection between imperatives and other TAM markers is reflected by the presence of these latter among command strategies in different languages. Eric W. Campbell demonstrates that, in Zenzontepec Chatino (Chapter 5), the potential mood is used in all other commands except for the second-person basic directives, which are realized by the imperative mood. This means that the potential occurs in non-addressee-oriented directives, prohibitives and all alternative command strategies. Campbell also draws attention to a possible morphological and historical connection between the imperative mood and the perfective aspect in this language. The two forms are marked in certain inflectional classes with a similar prefix, in other classes they are distinguished by the absence of a preposed nasal in the imperative prefix. This nasal is present in the perfect and is presumed to originate from a realis marker. The interplay of the imperative with irrealis mood, perfective aspect and future tense surfaces at several stages in the book.

When it comes to semantic distinctions specific to imperatives, they are motivated by the inherent property of commands to reflect the social structures

and the cultural meanings shared and maintained by the members of a speech community. Subdividing imperatives in delayed, immediate, plain, polite or familiar imperatives, as for example in Quechua (Adelaar, Chapter 2), Aguaruna (Overall, Chapter 3), Japanese (Jarkey, Chapter 8) and Nungen (Sarvasy, Chapter 11), is likely to be dependent on culture-specific values and norms concerning the relationship between the interlocutors. The fact that certain verb types are incompatible with imperatives can also be viewed as stemming from cultural meanings. Nerida Jarkey studies in detail the uses of imperatives in Japanese in terms of social (in)acceptability of commands. In contemporary Japanese, subject-honorific verbs tend to be at odds with plain imperative forms, since showing respect to someone who is socially above one's one position and issuing commands are actions that generally do not coincide in terms of social hierarchy. Another example of a limitation in imperative use are verbs expressing involuntary action: when it comes to language constructions, drawing the line between intentional beings and non-intentional entities, as well as determining the nature of this interface, is in the end a question of cultural values. The next section elaborates on some of the cultural and social issues examined in Aikhenvald & Dixon's book.

3 Cultural values and norms underlying commands

In the introductory chapter, Aikhenvald underlines the importance of being cautious when establishing links between the structure of a language and the ways of thinking of its speakers, as there is great risk of drawing too straightforward conclusions. Taking this into account, the book offers careful considerations on the complex relationship between language and culture, based on in-depth empirical analysis. The studies shed light on the reasons why mastering the different command strategies of a language is so crucial for successful communication and the ways in which language contact and cultural contact can change language use.

In many speech communities, issuing commands is conditioned by underlying social hierarchies based on age and gender, as well as social and emotional distance or familiarity. Azeb Amha analyzes a set of examples showing how plural forms are used to convey politeness in Wolaitta, even when talking to a single addressee (Chapter 14). In Japanese speech culture, issuing commands is in general particularly face-threatening and is therefore avoided when speaking to those above, in the vertical social dimension, or

to one's equals (Jarkey, Chapter 8). Japanese has a relatively large range of dedicated imperative forms, but it also employs a rich array of command strategies which make it possible for the speakers to avoid having to use plain imperatives. These strategies tend to lose their euphemistic qualities and gradually become more explicit. The choice of strategy ensues from the position of each interlocutor in the horizontal and vertical social dimensions (those above one's own status vs. those below, those who are in-group members vs. those with whom one does not easily identify with), the identity the speaker wishes to display and the relationship the speaker aims to foster with the interlocutor. Japanese command strategies are particularly sensitive to constraints concerning the ways in which socially acceptable gender identity is conveyed. According to Jarkey, certain authoritative ways of issuing commands are generally regarded as appropriate for men specifically. This can be challenging for example for women in professional positions of authority.

In many of the languages, there are special forms for interspecific commands. The speakers of Ashaninka Sapo (Mihás, Chapter 4) address commands to pets and other domesticated animals in order to call them to come or to chase them away. Some calls imitate the sounds produced by the animal (e. g. *ko ko ko* for chicken), others employ a diminutive form of the name of the species (e. g. *obisha obisha* < Spanish *oveja* 'sheep'). A long tradition of raising hunting dogs has led to a special category of dog commands in Nungon (Savary, Chapter 11). These commands are issued in order to direct the dog to alert the hunter to where game is located or to search at a certain location (*Ori horon!* 'Search on the ground!') or for a certain type of game. In Wolaitta, imperatives are used in directives addressed to oxen and cows (Amha, Chapter 14). In this case, the imperative is used in a singular form, regardless of the number of animals addressed.

It is possible that, in some languages, imperatives were originally used with reference to an even wider range of entities and that contacts with certain European cultures have influenced the way in which the environment is categorized. As a result, certain types of commands have become marginalized. In contemporary Quechua (Adelaar, Chapter 2; see also Aikhenvald, Chapter 1), impersonal verbs referring to weather conditions can be used only in third person imperative form. The author suggests that this limitation may not have existed in pre-Hispanic and pre-Christian era when the forces of nature could have been seen as potential addressees of commands. There are also some examples of Ashaninka Sapo and Karawari commands addressed to spiritual entities in Mihás' (Chapter 4) and Telban's (Chapter 13) papers.

The wide-spread idea that imperatives and commands are systematically and unequivocally face-threatening or interactionally problematic elements is also likely to result from contact between speech communities. Contrary to the situation in Japanese and many European languages, for Karawari speakers of Ambonwari village, commands indicate the close and positively direct relationship between human and non-human members of the community, as reported by Borut Telban (Chapter 13). Karawari commands are also an important part of language socialization through which children learn to participate effectively in the life of the speech community: in sharing, exchanges and cooperation.

In the concluding chapter of the book, Rosita Henry discusses the social aspect of directive speech acts from an anthropological point of view (Chapter 15). She calls attention to the whole of the social situation where commands are used. Instead of analyzing speech acts in terms of one-to-one relationships, it is necessary to take into account the other participants of the situation. Their presence is likely to influence the way in which the speaker addresses the interlocutor. The author illustrates the active role of listeners in a speech situation by presenting observations on the central position given to interpretation and understanding of talk among peoples of the Western Highlands in Papua New Guinea.

4 Conclusion

Aikhenvald & Dixon's book is a comprehensive panorama of imperative constructions and command strategies in the languages of the world. It brings to the fore the particularity of imperatives among verb paradigms and sentence types. The methodological regularity and thoroughness, reflected in the structure of each chapter, makes manifest the parameters of variation in imperatives and other command strategies and allows cross-linguistic comparison also with regard to specific issues in the semantics of commands. The different chapters furthermore undertake the important task of exploring the cultural meanings behind the use of directive expressions. In doing so, they open up perspectives to the effects of language contact in expression of commands and the ways in which language structures reflect speakers' means for positioning themselves within the society and categorizing their environment.

The book is of interest to all those seeking to gain insights into

cross-linguistic variation and patterning in verb paradigms and sentence types, language-specific interplay between verbal categories, and typology of grammatical constructions. It is also relevant reading to anyone concerned with the cultural and social aspects of grammatical constructions.

References

- Aikhenvald, Alexandra Y. 2010. *Imperatives and commands*. Oxford: Oxford University Press.
- Birjulin, Leonid A. & Xrakovskij, Viktor S. 2001. Imperative sentences: Theoretical problems. In Xrakovskij, Viktor S. (ed.), *Typology of imperative constructions*, 3–50. (LINCOM Studies in Theoretical Linguistics 9). Munich: LINCOM Europa.
- Miestamo, Matti & van der Auwera, Johan. 2007. Negative declaratives and negative imperatives: Similarities and differences. In Ammann, Andreas (ed.), *Linguistics festival: May 2006, Bremen*, 59–77. Bremen: Brockmeyer.
- van der Auwera, Johan. 2010. Prohibition: Constructions and markers. In Shu, Dingfang & Turner, Ken (eds.), *Contrasting meaning in languages of the East and West*, 443–475. Tübingen: Narr.
- van der Auwera, Johan & Dobrushina, Nina & Goussev, Valentin. 2003. A semantic map for imperative-hortatives. In Willems, Dominique & Defrancq, Bart & Coleman, Timothy & Noel, Dirk (eds.), *Contrastive analysis in language: Identifying linguistic units of comparison*, 44–68. Basingstoke: Palgrave Macmillan.
- Van Olmen, Daniël & Heinold, Simone. 2017. *Imperatives and directive strategies*. Amsterdam: John Benjamins.
- Xrakovskij, Viktor S. (ed.). 2001. *Typology of imperative constructions*. (LINCOM Studies in Theoretical Linguistics 9). Munich: LINCOM Europa.

Contact information:

Rea Peltola
Université de Caen Normandie
Laboratoire CRISCO
Esplanade de la Paix
CS 14032
14032 Caen cedex 5
France
e-mail: rea(dot)peltola(at)unicaen(dot)fr

Garcés-Conejos Blitvich, Pilar & Fernández-Amaya, Lucía & de la O Hernández-López, María (eds.). 2019. *Technology mediated service encounters*. Amsterdam: John Benjamins. Pp. 247.

Reviewed by Heli Tissari

1 Introduction

All readers of this journal are familiar with expectations and challenges related to using new technology. These are also the first topic discussed in the volume *Technology Mediated Service Encounters*. The context is a health centre that has adapted new technology to help nurses explain potty training to parents. In the second chapter, the context is a hospital which uses translation services over the phone. There then follow four chapters on call centres, discussing lengths of calls, interruptions to calls, protocols, politeness, the potential for negotiating deals, and the use of a second language in customer service. Lastly, the book contains three chapters on Internet interactions. The last section begins with online bargaining and moves through customer reviews to what happens in a conversation on a commercial Facebook wall.

All this is also explained in the introductory chapter of the book. Moreover, it briefly tells the readers how this volume relates to previous research, pointing out that the topic of service encounters is not new but that there are not yet so many publications on technology mediated service encounters. This is the research gap that the editors have decided to address. They consider it a strength that the chapters represent various methods and discuss several languages. Lastly, they say that research on technology mediated service encounters has “real life applications and impact” (p. 9).

In the following, I summarize the book section by section. The three sections cover social service, call centres, and e-service encounters.

2 Summarising the section on social service

De Wilde, Van Praet & Van Vaerenbergh have studied “Language discordance and technological facilitation in health care service encounters”. In practice, this means the use of a story-board app to explain potty training to parents. The authors want to know how the use of the app impacts service provider and customer satisfaction, whether it influences conversational sequencing and

eye gaze, and whether it helps communication. They compare visits where nurses use the app to visits where they do not, using exchange students to act as parents who are not proficient in Dutch. In brief, they notice that, while one of the studied nurses has learned to use the app very well, another hesitates about what to do, which leads to a different pattern of conversation and eye gazing. While both customers and service providers prefer to use the app, it is clear that more training is needed to make its use really efficient.

The second article similarly discusses problems which arise when participants are not wholly cognizant of how the situation is supposed to work. It is aptly titled “Context and pragmatic meaning in telephone interpreting”. Lázaro Gutiérrez & Cabrera Méndez focus on a telephone conversation during which an interpreter is supposed to help a doctor treat a patient in an emergency room. Fortunately, it is not about life and death, because it takes a long time before contact is established between the interpreter and the doctor, and both have identified their roles. Problems follow even thereafter because the patient does not immediately understand that he can speak his own language to the interpreter; neither does he want to fully comply when he gets the message. The authors consider it a challenge for the interpreters that they can be suddenly invited to various kinds of unpredictable situations, but, again, they suggest that at least part of the problems could be avoided through means of better planning and training.

3 Summarising the section on call centres

The name of Edmonds & Weatherall’s chapter “Managing verbal and embodied conduct in telephone-mediated service encounters” is slightly cryptic, as it refers to something that speakers on the phone have traditionally not been expected to see, the body. At the end of the chapter, they indeed speculate about what kind of phones the callers were using, surmising that speakers would have carried their mobile phones with them to check a meter reading, for example. This is what they refer to when they discuss ‘embodied conduct’. They are interested in what happens when a participant in the conversation has to move somewhere else to fetch information that is needed to continue the call. The outcome is that such activity may create a long silence in the middle of the call that can be avoided if the speaker keeps on explaining what s/he is doing. In the phone calls analysed by Edmonds & Weatherall, it was the clients who needed to multitask, and they may have had

landline telephones which could not be carried with them in order to keep the conversation going.

Hultgren's chapter on the globalization of politeness in call centres introduces a different topic – that of service personnel being trained to follow certain protocols. The gist is that if a similar protocol is used all over the globe, it is unlikely to always agree with the local context. In fact, there are differences on a much smaller scale. Hultgren analyses customer service interactions in Britain and Denmark and notices that the British staff follow the protocol conscientiously, while the Danes do not. The latter experience some of the protocol as unreasonable. A key example is that of providing and using personal names. A Danish informant explains that she feels uncomfortable if a strange person calls her by her first name; she in fact resorts to retaliating by using the customer's name. Interestingly, the Danish conversations are nevertheless no shorter than the British ones.

Revenge is also potentially an issue in Márquez Reiter's article "Navigating commercial constraints in a service call". She analyses a conversation where a salesperson fails to profit from a clever client who attempts to turn the occasion to her advantage. In the beginning of the conversation it seems that the salesperson has the upper hand and that the client is relatively unknowledgeable and innocent, but as the call proceeds, she shows that she has some aces up her sleeve. At some point, her attempt to close the conversation is nevertheless turned down by the salesperson who called her in the first place. Eventually, neither participant gains from this particular conversation.

While Márquez Reiter's negotiators are very skilful speakers of Spanish, Mugford's chapter on second-language English speakers working in a call centre in Mexico that serves U.S. clients introduces customers' doubts concerning language proficiency. It discusses problems encountered by the call centre staff and the ways they report solving them. The key terms *conversationalisation*, *(im)politeness* and *discursive practices* appear in the title of the study. It can be assumed that the Mexican staff would have more conversational resources at their disposal if they used their first language; however, they have developed clever discursive practices to avoid and solve conflicts. If there is a problem, it tends to be the client who is impolite, while the call-centre agents have to remain polite to the extent that they may not have the right to close a conversation even if it gets very heated.

I would like to commend Mugford for his choice of topic as it acknowledges the value of Mexican call center workers that is otherwise

overlooked, violating their quintessential right to be respected for their contributions. This in essence allows the professionalism of his subjects as well as their unique perspective to become apparent. This reminds me of an article by Chouliaraki (2011) where she underlines that solidarity should consist of bringing to the fore the people who actually suffer from a wrong, and their experience and reports, rather than of the commentator's views from the outside.

4 Summarising the section on e-service encounters

Placencia's chapter deals with refusals of offers on a website called Mercado Libre Ecuador. More specifically, its focus is twofold: sometimes the prospective buyers bargain, and sometimes they barter. Both are somewhat surprising as these behaviours are not encouraged by the website management. Placencia discovers that sellers respond differently if potential buyers offer a lower-than-suggested price than if they offer something other than money in return for the sales item. A buyer trying to barter is more likely to receive an explicit refusal than one trying to bargain, whose request is more likely to be met by an implicit refusal, such as a counter-offer. Placencia is also interested in the extent to which sellers' responses include supportive moves, such as affiliative address terms and apologies. They are most likely to contain greetings and least likely to include a signature.

Ren's study on intensification in online consumer reviews takes us to another continent, China. He has counted occurrences of the following in Kindle book reviews: preceding intensifiers, postpositional intensifiers, expletive/taboo words, metaphors, repetitions and emphasis by means of punctuation. While explaining these matters he also explains expressions and habits that may appear strange to writers operating with other languages. One of these is the trend to write the same sentence three times to emphasize it, a habit which has become viral in China since 2015 although commentators disagree about why. Ren observes that the choice of intensifier depends on whether the review is negative or positive, although most of the frequent intensifiers look rather "innocent" on the surface – words that can be translated as 'very' and 'too', for example.

In the last chapter of the book, Bou-Franch analyses "relational practices on commercial Facebook wall interactions". She has collected a reference corpus of 10 wall interactions from the Facebook page of an American

store and chosen one to focus on. This interaction mainly consists of customers' responses to a post by the store, but also contains some comments by representatives of the store. The majority of the customers' posts are supportive of the store and its products, but a number of them are unsupportive, creating a dynamic that Bou-Franch is interested in, especially since the unsupportive posts begin with the appearance of a criticism not related to the particular product that is advertised. She thus reports what happens when a customer begins to criticize the store for an issue unrelated to the advertisement and is backed by like-minded people.

5 Discussion

The editors of this volume have clearly chosen a topic that interests many linguists beyond its authors. Much is happening right now. For example, Chatwin & McEvoy (2019) describe how they plan to use recordings of video conferences from a carer support initiative called Empowered Carers to analyse conversations between customers and support workers. An equally recent article by Bloch & Leydon (2019) reviews helpline practices and points out practical applications of conversation analysis. More generally, service encounters have also been of interest, for example, to Choi et al. (2019), who have compared literal and figurative language styles in interactions between humans, humans and robots, and humans and service kiosks.

Considering that this kind of research seems to be mushrooming at present, it is important for linguists interested in it to read this volume. It can be used as an introduction to technology mediated service encounters, or as a valuable series of current snapshots.

However, since further research is going on at the moment, there will probably soon be a need to publish more volumes on the topic. One can, for example, review and categorise new studies to see how much research has been conducted on a more specific topic such as healthcare apps and which specific topics are the most popular. It might then be reasonable within a couple of years to edit further volumes on more specific topics. One such topic could be healthcare; another could be problem navigation; and so on. For example, one could compare the negotiation of commercial constraints (as discussed by Marguez Reiter in this volume) with skilful suppression of complaints (Kevoe-Feldman 2018). More popularized approaches can also be envisioned: someone could share insights concerning technology mediated

service encounters in a book directed at laypeople.

To continue with laypeople in the sense of people who are not professional linguists, I discussed many of the ideas presented in this volume with a friend who works in the public sector in Finland, answering phone calls. She could relate to the contents of the book in various ways. For example, she told me that she and colleagues always provide the name of their employer and their own name when they respond to a call, but otherwise they do not follow a fixed protocol. She also told me that an experienced adviser is able to listen to the customer so as to tailor the conversation to that specific person's needs and to talk to them in a way that they can understand. Such an adviser will explain to the customer if s/he needs time to find and process information, so that no awkward silences occur on the phone. I feel that my conversation with her was ripe with potential ideas for further studies.

It thus appears that the authors of the book are right: it is useful for researchers to interact with people who perform technology mediated service encounters. As the editors of the book suggest, this can certainly have “real life applications and impact” (p. 9). They and their chosen authors have succeeded in choosing topics that have real life relevance. What therefore strikes me as a reader is that it is indeed important that linguists who study such practical issues discuss their findings with people who can profit from them, be it service providers or recipients – even the general public.

References

- Bloch, Steven & Leydon, Geraldine. 2019. Conversation analysis and telephone helplines for health and illness: A narrative review. *Research on Language and Social Interaction* 52(3). 193–211.
- Chatwin, John & McEvoy, Phil. 2019. Conversation Analysis (CA) as a tool for exploring interaction in an online video-conferencing based support service. *Journal of Enabling Technologies* 13(3). 150–157.
- Choi, Sungwoo & Liu, Stephanie Q. & Mattila, Anna S. 2019. “How may i [sic] help you?” Says a robot: Examining language styles in the service encounter. *International Journal of Hospitality Management* 82. 32–38.
- Chouliaraki, Lilie. 2011. ‘Improper distance’: Towards a critical account of solidarity as irony. *International Journal of Cultural Studies* 14(4). 363–381.
- Kevoe-Feldman, Heidi. 2018. The interactional work of suppressing complaints in customer service encounters. *Journal of Pragmatics* 123. 102–112.

Contact information:

Heli Tissari
Department of English
Stockholm University
e-mail: heli(dot)tissari(at)helsinki(dot)fi

Contents:

- 5 Reviewers of *SKY JoL* 32 (2019)
- 7 Saleem Abdelhady
✚ A sociolinguistic study of the use of the definite article (*ʔal*) in the Jordan Valley
- 43 Marta Coll-Florit & Salvador Climent
✚ A new methodology for conceptual metaphor detection and formulation in corpora: A case study on a mental health corpus
- 75 Maarja-Liisa Pilvik
✚ Assessing the productivity of the Estonian deverbal suffix *-mine* in five registers of Estonian
- 105 Hanna Pook
✚ The pronoun *kes* 'who' and its referent's animacy in Estonian dialects
- 145 Ulla Vanhatalo & Heli Tissari & Taru Lilja & Kimmo Vehkalahti & Mari Siirainen
✚ "Something bad can now happen to me here": Meaning components of emotion words

Book Reviews:

- 181 Abdelkader Fassi Fehri. 2018. Constructing feminine to mean: Gender, number, numeral, and quantifier extensions in Arabic
- 187 Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.). 2017. *Commands: A cross-linguistic typology: Explorations in linguistic typology*
- 195 Pilar Garcés-Conejos Blitvich & Lucía Fernández-Amaya & María de la O Hernández-López (eds.). 2019. *Technology mediated service encounters*

Publisher:

Suomen kielitieteellinen yhdistys
Språkvetenskapliga Föreningen i Finland
The Linguistic Association of Finland