Reading Dolgans through their language: approaching indigenous Dolgan culture by collecting words. An exploration into a quantitative methodology for comparative semantic analysis.

Reijnaers, Damiaan

Universiteit
Leiden

Reading Dolgans through their language: approaching indigenous Dolgan culture by collecting words.

An exploration into a quantitative methodology for comparative semantic analysis.

A thesis presented for the degree of Master of Arts
in Russian and Eurasian studies

University Leiden, July 1, 2022; Faculty of Humanities

Student: Damiaan J.W. Reijnaers
First supervisor: **prof. dr. E.L.J. Fortuin**
Second supervisor: **prof. dr. J. Schaeken**

Abstract

This thesis builds on the idea that subtle, culturally induced differences in semantic meaning remain between translation equivalent words across different languages. This study further argues that these differences in meaning may be approached through the examination of the linguistic contexts within which these words occur. Consequently, this work provides a quantitative methodology for highlighting relevant areas in which such cultural differences may be reflected. The method is based on intuition derived from several existing, structuralist methods and works primarily by comparing the frequency of hypernyms of nouns that appear in the neighborhood of an examined word. This thesis focuses on the indigenous Dolgan language as a case study; one that is purposely exploratory in nature. This minority language poses the research with the additional challenge of working with a small-sized language corpus for computational purposes: it demands a 'rough' look at data to act as a means, instead of being a limitation. Overall, the results indicate that culturally determined differences between words exist to a measurable degree, despite the unavailability of an adequately sized dataset. Although the results provide insufficient guidance for drawing anthropological conclusions, the findings reassert that cultural knowledge is encoded within language and reiterate the need to preserve endangered indigenous languages.

Contents

List of Figures

List of Tables

1. Introduction

> *For peoples without a script, which include the Dolgans, the linguistic expression of a world view is extremely important, as the complex linguistic constructions and compositions are the fruits of a creative path that preserved the core value of the whole of the culture.*
>
> – Danilova (2017, p. 85)

When observing two different languages, one might notice and focus on words that refer to the same physical entities (Morris, 1938, p. 10); thereby neglecting the cultural dimension of language, which might cause subtle differences in the conceptual interpretation of seemingly equivalent words. For example, the semantic meaning of common, everyday, words in English, such as 'cup' or 'potato', differ somewhat from its translation equivalents in Polish (Wierzbicka, 1985, p. 4).

The Russian Federation is home to more than a hundred different languages (Moseley & Nicolas, 2010, pp. 42, 48–52, 54, 56, 57). Many of these languages are characteristic of various cultures, especially the languages that are indigenous to the Russian Federation (Diachkova, 2001, p. 223). The country therefore forms an attractive testbed for linguistic research into cross-cultural semantic differences, as certain non-cultural factors are held more constant relative to cross-border comparisons (these factors include, *e.g.*, societal aspects, or, in some cases, geographical influences). In this thesis, I attempt to defend the view that the semantic differences that remain between the Russian language and indigenous languages are overwhelmingly the result of cultural differences that can be explained through anthropological methodology.

Drawing inspiration from the theory of distributional semantics (Sahlgren, 2008), which builds on the hypothesis that 'words acquire semantic meaning by the linguistic contexts in which they appear,' this thesis further maintains that *cross-cultural* differences between words can likewise be characterised by a comparative analysis of the surrounding words with which they occur. The intuition behind this assertion follows the rationale that cultural behaviour is inherently mirrored in language describing this behaviour—the words used to describe a cultural event are characteristic of the corresponding culture by the occurrence of their combination and the mere presence of these words themselves. For example, the common use of dogs for transport could be considered culturally dependent and attributed to certain indigenous cultures within the Russian Federation. Therefore, it could be reasonably assumed that the word 'dog' would co-occur more frequently with words indicating transport (*e.g.*, 'sled'); by contrast, the perception of dogs as pets would be more regularly echoed by Russian speakers. The linguistic *context* within which certain words appear can thus reflect the cultural habits of the speaker.

An important prerequisite of such contextual analysis is the availability of an adequately sized dataset of linguistic productions in the languages of interest. Such datasets

are called language corpora and form the backbone of corpus linguistics. Corpus linguistics is a methodology that studies language through analyses on the basis of large collections of authentic examples of language use. In a way, the suggestion by John McHardy Sinclair—one of the pioneers of corpus linguistics—that meaning arises "through several words in a sequence" (Bennett, 2010, pp. 8–9) can be seen as an early version of the distributional hypothesis from which this thesis takes its inspiration.[1] For Sinclair, this idea was the foundation on which he helped develop corpus linguistics in the first place (Bennett, 2010, p. 2). The number of studies that utilise corpus linguistics is increasing, especially for making quantitative statements (Joseph, 2008, p. 687). However, while such data for the Russian language are readily available (Zakharov, 2013), approaches of this kind carry with them limitations in terms of applicability to languages without an availability of large-scale textual corpora; this especially holds for indigenous languages.

Another complicating factor is the methodology involved with the subsequent quantitative analysis of the contexts extracted from the corpora and the semantic differences between languages they might indicate. Although quantitative studies conducted via corpus-based methods have recently gained momentum in linguistics, quantitative methods in general have formerly often been met with resistance among linguists and have therefore not been thoroughly developed, as is the case in other fields (Gries, 2013, pp. 4–5). Moreover, languages with a small population of native speakers generally receive less attention from the linguistic research community than widely spoken languages. This especially holds for research approached by statistical or computational methods. However, quantitative analyses are useful as they provide a solid numerical interpretation which could be extended to future work involving different languages or research angles. A further advantage of employing a quantitative approach is that it allows for a visualisation of the observed differences, as the (interpretable) contexts can be defined as the dimensions of a geometric space in which the to-be-investigated concepts are drawn, based on their respective frequency counts for every observed context. An intuitive example is illustrated in figure 1. Ultimately, linguistics is a research field yielding almost exclusively practical implications. Therefore, a quantitative analysis functions merely to finally inform *qualitative* judgements on the conceptual interpretation of the target words (Strauss & Corbin, 1998, p. 34). In this thesis, such qualitative evaluation is further aided and directed by scholarly anthropological literature.

[1] As with many highly influential ideas in science, several different theories exist with respect to the distributional hypothesis; not all of which are in favour of using corpus-based methods for applying it (Montes, 2021, p. 2). As the distributional hypothesis (regardless of in which form or through which method) merely acts as a source of inspiration for my reasoning, and as the base of the theory requires only a simple explanation (which is already given), I consider an extended elaboration on the hypothesis itself inappropriate in this instance. A recent overview of works related to the hypothesis can be found in the PhD thesis by Montes (2021, pp. 1–10), which was supervised by one of the authors frequently cited in this work.

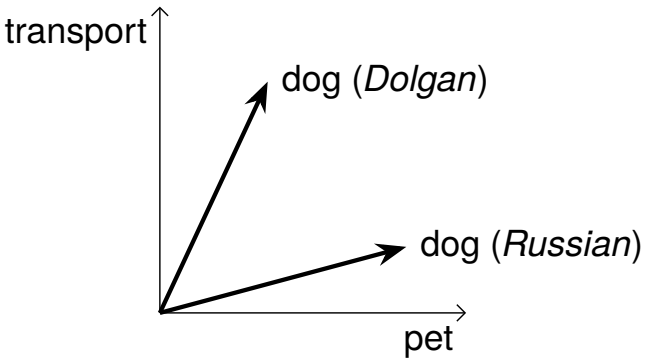


Figure 1. An example of a context space. If the word 'dog' in Russian would frequently appear in sentences, such as *"My friend owns such a cute dog."*; whereas in Dolgan, the word would tend to appear in sentences, such as *"He sat down onto his dog sled."*; assuming a generalisation into only two types of contexts, the resulting geometric space based on these contexts could look like this illustrated graph.

My thesis adopts a case study approach to examine the differences in contextual usage of a set of frequently used nouns between the Russian language and the Dolgan language. Considering that my thesis concerns comparisons across—possibly very differently structured—languages, focusing solely on nouns makes sense as they appear universally in all natural languages (Dixon, 2014, pp. 39–41), making them a fit for any language pair.[2] Moreover, nouns are the most basic elements in a language (Gentner, 1982, p. 301), and so they are the 'prime' holders of meaning. The choice of indigenous language is motivated by the relatively large amount of available annotated data, compared to data on other indigenous languages spoken within the Russian Federation.

Dolgan people are located in Russia's *krajnyj sever*, or 'the extreme North', in the Taimyr region. The language was spoken by only 1,054 people in 2010.[3] The Dolgan language is a Turkic language and, unlike Russian, not a descendant of the Indo-European family of languages. As these languages themselves thus developed further away from each other, consequently, the etymological ancestry of these languages' frequently used nouns are also likely further apart. Intuitively, therefore, it only seems plausible that (subtle) differences in meaning have likely emerged between these core nouns, when envisaging that both languages 'started with a blank slate' on which they could, relatively independently from each other, 'draw the meaning' of such nouns (which apparently emerged in both

[2]The statement that nouns appear universally in all languages might spark controversy among some scholars, as certain languages' nouns behave similarly to verbs (Darkgamma, 2014). However, the cited works by Dixon (2014, p. 41) and Gentner (1982, p. 327) make explicit that a distinction between verbs and nouns always exists (and that nouns therefore are likely linguistic universals).

[3]Federal'naja služba gosudarstvennoj statistiki, *Itogi Vserossijskoj perepisi naselenija 2010 goda v otnoenii demografičeskix i social'no-ekonomičeskix xarakteristik otdel'nyx nacional'nostej*. Retrieved from: https://web.archive.org/web/20220105150143/https://www.gks.ru/free_doc/new_site/perepis2010/croc/results2.html (5th of January, 2022).

languages). This linguistic trait makes Dolgan thus even more attractive as a case study, in comparative relation to the Russian language.

My study is largely based on the INEL Dolgan Corpus, which, at the time of writing, contained about 10 hours of annotated speech and about 30,000 words of folkloristic text (Däbritz, Kudryakova, & Stapert, 2019). The Russian samples are provided by a small version of the Russian National Corpus (RNC) that contains about 1,000,000 words (Furniss, 2013, p. 200). Importantly, all texts within the Dolgan corpus have been manually translated into Russian, which allows a comparison of nouns from both languages to be made.

The study of *how cultural factors contribute to meaning* is an important aspect within the field of anthropological linguistics as findings in this realm (of ethnosemantics) could provide a useful anthropological account of the environment in which a language is spoken. My study provides an initial exploration to what extent such anthropological implications follow from the explicit linguistic context of meaningful concepts (*i.e.*, the surrounding words in a sentence) and whether their projection onto a geometric space defined by precisely these contexts can function as an adequate tool for finding such implications. Therefore, this study does not merely attempt to identify differences in terms of the semantic meaning of everyday words across Russian and Dolgan; it further aims to contribute to the field of (linguistic) anthropology by providing a computational means for inferring cultural traits. The central research question that I attempt to answer in this paper can be formulated as follows: *how can differences in semantic meaning between translation equivalent words in Dolgan and Russian be characterised by the linguistic context in which these words occur?*

This thesis has been organised in the following way. In section 2, the paper first gives a brief overview of the culture and the traditions of Dolgan people. It will then naturally proceed by defining the concepts of interest based on the presented cultural findings and the contents of the INEL Dolgan Corpus in section 3. This section is further concerned with the methodology used for this study with regard to the processing of the corpus data (in section 3.1) and the quantitative approach taken (section 3.2). The 4th section presents the findings of the research which will finally be discussed in section 5.

2. Dolgan people

Dolgan culture sprang from the cultures of Yakuts, Russians, Evenks (Kistova et al., 2019, p. 793), and other indigenous groups (Alekseevič & Efremov, 2008, p. 427). Elements from their ancestors' cultures, as well as from other cultures (such as Nganasan and European influences) can also be found in Dolgan folklore (Alekseevič & Efremov, 2008, pp. 431–432, 434, 437). Until recently, Dolgan people didn't even refer to themselves as 'Dolgan' (Alekseevič & Efremov, 2008, p. 427), which makes sense as the group is relatively young (Bettu, 2011, p. 287), and as the name has been invented by Russians (Alekseevič & Efremov, 2008, p. 427). As Danilova (2017, p. 85) puts it succintly: "The intertwining of ethnogenetic lines and waves of migrations and long term neighbor-ship in contiguous territories led to the emergence of the cultural identity of Dolgan people."[4] Moreover, the construction of the Dolgan *written language* also only took place during the later decades of the Soviet Union (Siegl & Rießler, 2015, p. 209), which allowed the researchers cited in this section to preserve Dolgan folklore in its original forms (Bugaeva, 2013, p. 83).

The subsections that follow provide a thematic overview of historical and modern features of Dolgan culture that could be reflected in the Dolgan language, as the experiments of this study attempt to discover. This section aims to present a *general picture* of Dolgan culture—that I by no means claim exhaustive—and provides the basis for the theory discussed in section 3.2.1.

Importantly, the findings in this section identify a clear distinction between Dolgans' ancient and modern ways of life.

2.1. Lifestyle

Traditionally, as a direct consequence of their nomadic lifestyle (Popov, 1934, p. 123), the primary occupations of Dolgan people were reindeer herding, fishing, and hunting (Zamarayeva, Kistova, Pimenova, Reznikova, and Seredkina (2015, p. 227); Bettu (2011, p. 287)), which were pursued by both men and women (Popov, 1934, p. 127). Various cultural norms are connected to these occupations: *e.g.*, the 'rule' that a person may never throw away parts of a reindeer or fish, or that a hunter should refrain from hunting too many prey (Bettu, 2011, p. 291).

To this day, Dolgan people are still engaged in hunting and fishing activities, especially in more northerly villages as opposed to southern settlements (Zamarayeva et al., 2015, p. 227), although Bettu (2011, p. 288) also observed fishing and hunting activity by

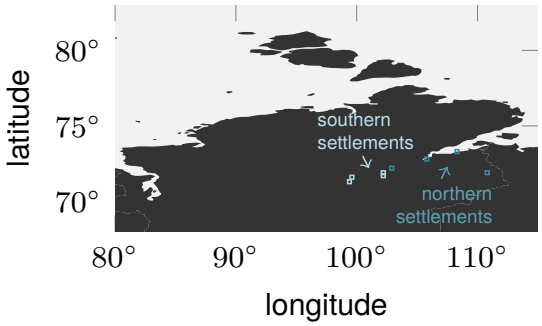
[4] Original fragment: *'Perepletenie etnogenetieskix linij i voln migracij, dlitel'noe sosedstvo na smenyx territorijax opredelilo vozniknovenie samobytnoj kul'tury dolgan, sxodnoj s tradicijami "kontaktnyx etnosov".'*

elderly people in a southern village. In current times, reindeer herding and fishing became more or less official jobs with salaries paid for in rubles, as observed by Davydov (2016, p. 76) (indeed in northern locations). In general, reindeer herding now increasingly bears a symbolic nature (Zamarayeva et al., 2015, p. 231) pursued mainly for the purpose of sustaining a cultural identity (Zamarayeva et al., 2015, p. 227). Moreover, nowadays, reindeer races only take place on 'den' olenevoda' 'reindeer herders' day' (Davydov, 2016, p. 80); a holiday that is celebrated by many northern indigenous peoples and involves several activities such as markets and contests (Momzikova, 2013, p. 227). Dolgans also hunt for geese and partridge (Davydov, 2016, p. 79), and, since the Russians showed their economic value, for fur-bearing animals (Popov, 1934, p. 128). Men are responsible for fabricating the nets used for hunting (Popov, 1934, p. 127). The same nets are often used for different animals, although rifles are sometimes also used to hunt geese (Davydov, 2016, p. 79).

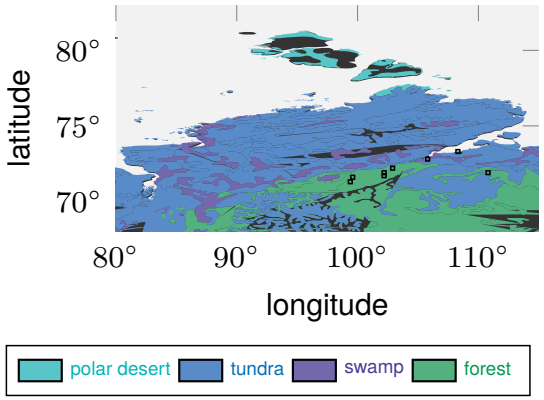
2.2. Habitation

Dolgan people live on the Taimyr peninsula (see figure 2a). Their habitat is almost completely surrounded by tundra and forest (figure 2b). In current times, Dolgan people live in modern cottages in villages that contain administrative facilities, libraries, kindergartens, schools, and post offices; their homes have flat-screen televisions and access to Wi-Fi (Davydov, 2016, p. 67, 68, 72, 74). Many Dolgans have emigrated to cities wherein which they primarily speak in Russian to the detriment of their native language (Bettu, 2011, pp. 288–289). Traditionally, Dolgans pursue a nomadic lifestyle in summer while remaining sedentary in winter (Popov, 1934, pp. 123–124). Therefore, traditionally, in winter, Dolgans live with their families in separate nomadic tents ('čumy'), while, in summer, one tent is shared by about three families to save cargo when trekking to another spot (Popov, 1934, pp. 123–124). In autumn, when groups are formed for the summer, to gain a higher chance of survival against starvation, Dolgan people prefer to join groups with rich fellow Dolgans, as everyone is obliged to share food among fellow group members (a family that has about twenty reindeer is considered 'rich') (Popov, 1934, pp. 125–126). In summer, some families stay close to rivers and lakes in order to stock up on fish for the winter (Popov, 1934, p. 125). During the cold winter months, families may also stay in forests (Popov, 1934, p. 123).

In modern times, this nomadic tradition is lost and nomadic tundra routes are forgotten (Zamarayeva et al., 2015, p. 227). However, although dog sleds have largely been replaced by snowmobiles, traveling using dog sleds remains popular for short distance travel (*e.g.*, during hunting or to go fishing), especially for people without a snowmobile (Davydov, 2016, pp. 82–84). Many Dolgans now make use of GPS services, and Dolgans who own snowmobiles have to be cost-conscious about fuel (Davydov, 2016, p. 78, 86).



(a) Dolgan villages in Tajmyrskij Dolgano-Neneckij rajon, Krasnojarskij kraj, Russia. Settlements are divided into 'northern' and 'southern' settlements, following Zamarayeva et al. (2015, p. 227).



(b) Vegetation types of and around the Taimyr peninsula (Stolbovoi & McCallum, 2002).

Figure 2. Geographical habitation of Dolgan people.

In earlier times, Dolgans used nine dogs to ride 70 to 80 kilometres per day, by which they could move 700 kilograms of cargo; now, only three to five dogs are used (Davydov, 2016, p. 84).

2.3. Society

Already before the Soviet revolution, princes were elected as the head of Dolgan clans and carried out the will of the Tsar. This included collecting taxes and performing judicial functions (Popov, 1934, p. 135).

Dolgans used to behave largely collectively: *e.g.*, by sharing tools with each other without any consideration, or by sharing the cost of dowries (Popov, 1934, pp. 128–129, 133). Nowadays, individualism is gaining momentum (Zamarayeva et al., 2015, pp. 127–128). However, while Davydov (2016, p. 69) observes that many Dolgans now commercially trade with each other, Davydov (2016, p. 78) also observes that fishermen still share their first catch following cultural habits (Popov, 1934, p. 127). Moreover, Dolgans collectively decided to ban the sale of alcohol from their stores (Davydov, 2016, p. 71), and warn others of dangerous fishing weather using signs (Davydov, 2016, p. 77).

Future partners meet each other during dancing, after which they send each other gifts. Marrying is allowed with individuals further than three generations (although exceptions are sometimes made) and sex before marriage is generally accepted (Popov, 1934, p. 131).

2.4. Religion

As a consequence of Dolgans' mixed ancestry, the religious views of Dolgans likewise have mixed origins. For example, Russian Orthodox features, such as icons, are overtly present within their surroundings (Kistova et al., 2019, pp. 793–794, 805). Popov (1934,

p. 131) mentions the celebration of Christian holidays, such as Christmas. Other influences stem from totemism or animism (Kistova et al. (2019, p. 794); Alekseevič and Efremov (2008, p. 436); Popov (1934, p. 134)). In Dolgan folklore, animal roles are common, *e.g.*, the character of a 'deceptive fox' (Alekseevič & Efremov, 2008, p. 433). Another example is the depiction of a horse as 'the progenitor of life' (Kistova et al., 2019, p. 795). Terminology related to hunting and reindeer herding often come up in folklore that tells about the creation of the world (Danilova, 2017, p. 87).

However, the most important ideas are those of shamanism (Kistova et al. (2019, p. 795, 804–805); Alekseevič and Efremov (2008, pp. 436–437); Popov (1934, p. 134)), many of which have been influenced by Yakutian culture (Popov, 1981, p. 254). A shaman protects one or more individuals (Popov, 1934, p. 135). There exist three different 'types' of shamans (formerly seven), depending on the degree of skill of the shaman: a 'weak' shaman, an 'average' shaman, and a 'big' shaman (Popov, 1981, pp. 254–255). Each shaman owns a tree from which the shaman's skill can be read (big shamans own three), *i.e.*, if a shaman is strong, the tree will grow lots of branches, while, if a shaman is weak, the tree will also look weak (Popov, 1934, p. 135). Different types of ceremonies do also exist and are distinguished by their level of complexity: '**algɨs**' (prayerful appeals and requests to friendly and unfriendly spirits; used for 'easier' problems), and '**kɨːrɨː**' (a more complex ritual which involves special clothing and attributes, and might take multiple days; used exclusively for very serious cases) (Popov, 1981, pp. 255–256). Shamans may be consulted for numerous issues, *e.g.*, during wars, famines, or enemy attacks, or, *e.g.*, for threatening ill people, or asking for help in love (Popov, 1981, pp. 258–259). In everyday life, shamans participate in fishing and hunting activities like all other members of the community (Popov, 1934, p. 135), as the income from performing rituals is not sufficient to sustain a living (Popov, 1981, p. 260).

Dolgan religion is furthermore centered around *world tree* symbolism, which implies the existence and interconnection of 'three layers of life': an *underworld*, a *middle world* (or terrestrial world) and an *upper world* (or heaven), as described in detail by Danilova (2017, pp. 88–96). For this research, it is mostly relevant to note that the motifs in folklore conveying such symbolism include mountains, water, and forests (Danilova, 2017, pp. 90–92).

3. Methodology

The cultural background of Dolgan people could have a (subtle) influence on the Dolgans' semantic interpretation of certain words, which might be revealed upon a comparison with a language spoken in relative geographic proximity by people of a different culture, *e.g.*, the Russian language. Now that the cultural context of the Dolgan language has been summarised in the previous section, this section will focus on the procedures for actually comparing the semantic meaning of words between the languages. These comparisons are, in turn, guided and tested by the cultural knowledge presented in the previous section. As these comparisons are based on corpus data, the amount of possible word comparisons is moreover limited by the size of the corpora.

In order to gain a first insight into the type of data available, this section begins with an initial exploration of the INEL Dolgan Corpus and an explanation of the means for parsing it. As the methodology of this work primarily aims to shed light on the cultures of indigenous, minority languages (*e.g.*, Dolgan), an elaboration on the contents of the reduced version of the Russian National Corpus is omitted. However, for the purpose of reproducibility, the procedure for parsing the 'mini'-RNC will still be described. This section will then go on by laying out the theoretical dimensions of the desired methodology. This theoretical framework will naturally lead to a short overview of the relevant, currently available techniques for comparing the meaning of words, the combination of which will constitute the methodology of this research that will be presented directly afterwards. Finally, the theoretical background will additionally permit a closer analysis of the data, using which the words of interest to be compared will be specified.

This section has been divided into three parts. The first part deals with the corpus data. The second part deals with the specifics of conducting the comparisons and is by itself broken down into three 'subsubsections,' respectively concerned with the methodological background, existing methodology, and the newly proposed method. Based on all the foregoing, the third subsection will define the comparative experiments.

3.1. Corpora of the Dolgan and Russian language

Däbritz et al. (2019) provide freely available access to all the Dolgan corpus data and complementary documentation: see my bibliography for the digital object identifier and the corresponding URL. All of the work presented here is based on version `1.0` of the corpus. In this thesis, I follow the INEL conventions for transliterating Dolgan language, as described by Arkhipov (2020, pp. 4–9). As Dolgan was initially a language without a written tradition (see §2), many of the texts (in)directly stem from oral narratives.

Compared to the 76,912 tokens[5] of the Dolgan corpus, the Russian National Corpus, the size of more than a billion[6] tokens, is much larger. However, searching the RNC requires querying through an online interface, which becomes very labour-intensive if all words have to be manually collected. For that reason, I reached out to the organisation that operates the corpus to obtain an offline, smaller version of the data, amounting to about a million tokens. Despite its highly reduced size, this 'mini'-RNC still contains a considerably larger number of tokens than the Dolgan corpus; only a smaller number of samples would have been problematic, as that would have made a comparison entirely infeasible.

This section first gives an overview of the features of the corpus that are relevant to my research. Additionally, it provides a procedure for working with these features; one that is deliberately concise as to avoid lengthy description of technical details that are out of the scope of this paper.

3.1.1. Procedure for parsing the corpus data

The texts within both corpora have been subdivided into various categories. Each category contain several texts, which in turn contain multiple sentences of words. The 'one-million-tokens'-version of the Russian National Corpus features five categories: online blog posts, fiction novels, newspaper texts, scientific publications, and speech (*i.e.*, conversations). I have included all these categories in my experiments, except for the category consisting of scholarly articles. This choice was based on the intuition that science ought to be as neutrally written as possible, while the topics of the texts themselves also leave less room for cultural influences overt in the contexts of nouns. Although it is very well conceivable that cultural traits are reflected in *any* Russian text, incorporating scientific articles might only complicate this paper's experimental exploration.

The texts of the INEL Dolgan Corpus are categorised either as folklore, narrative, conversation, song, or translation. The latter category concerns texts translated from Russian into Dolgan and is for obvious reasons not included in my experiments. As songs contribute only 99 words to the corpus, texts within this category are also not taken into account. The former three categories are all included in the experiments and amount to the aforementioned total number of 76,912 tokens.[7]

[5] A 'token' is the base unit of a language corpus; it is usually a word. As the base units of the INEL Dolgan Corpus and the Russian National Corpus are indeed words, the terms 'word' and 'token' can be viewed as interchangeable in this thesis. As 'token' is more standard when writing about corpora in a scholarly context, I use the term when I directly refer to a corpus

[6] Nacional'nyj korpus russkogo jazyka, *Statistika korpusa*. Retrieved from: https://ruscorpora.ru/stats (26th of June, 2022).

[7] These data are derived from charts provided by INEL. See: https://inel.corpora.uni-hamburg.de/charts/dolgan-1.0-charts.html (Retrieved 2nd of May, 2022).

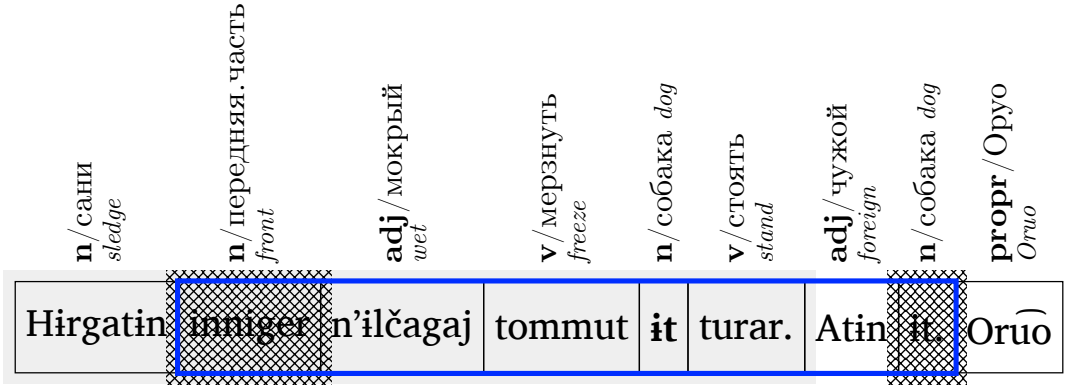


Figure 3. Example of a sliding window through the INEL Dolgan Corpus. An excerpt of a text within the corpus is shown at the bottom. A piece of this excerpt, marked in dark gray, is magnified at the top of the figure. The part-of-speech tags (abbreviated as n(oun), v(erb), adj(ective), or prop(e)r (noun)), and Russian and English translation annotations are shown above each separate word of the magnified piece of text. In this example, ɨt (in bold) is the target word. Nouns in the neighborhood of the target word are the context words (marked with a crosshatch pattern). The blue box (from inniger to ɨt) represents a sliding window of size 3 (three words to the left, and three words to the right). The light gray strip (from hɨrgatɨn to turar) marks the sentence of the target word (bold-face ɨt).

Although both corpora provide their data in different formats, all data files are ultimately organised using XML and can therefore be processed with only limited knowledge of programming. Crucially, all data contain manually annotated part-of-speech tags, and a Russian (translation of the) stem in the case of nouns. For this particular study, additional syntactic annotations are not needed.

As stated in the introduction, the idea behind the methodology is to approximate the meaning of a noun *by the other nouns that surround it*. To rephrase it, the idea is to define a 'target word' by its 'context words'. This can be intuitively visualised as a 'window' sliding through the corpus, with the target word at its center. All nouns that appear within the window are the context words. This intuition is illustrated in figure 3. The target word itself is not a context word. However, different occurrences of the same target word might count as context words as long as the words appear within the window centered around the target word. As additionally shown in the figure, some single words in Dolgan translate into multiple words in Russian (*i.e.*, Dolgan 'inniger' translates into Russian 'perednjaja čast'). In these cases, the entire translation is regarded as a single context word.

Normally, the larger the size of the sliding window (typically set at 4 or 5), the lower the significance of the experiments (Lindquist, 2009, p. 73). For my experiments, I test

different windows. Apart from experimenting with the size, I additionally experiment with limiting the window's scope by the sentence boundaries of the target word. In figure 3, this would entail that the second occurrence of '**ɨt**' would not count as a context word, as it is outside of the sentence of the target word (also '**ɨt**'). The scope of the sliding window is always limited by the boundaries of the corresponding text, as texts generally have been independently collected and were produced by different speakers on different topics (*i.e.*, a noun located within another text than that of the target word, despite being part of the same corpus, can never act as a context word). A number of texts within the INEL Dolgan Corpus feature multiple speakers and are consequently annotated with the speakers' corresponding identities. Evidently, this typically occurs when speakers engaged in a conversation on a certain topic. In my thesis, I also experiment with allowing the scope of the window to pertain to words uttered by a different person than the speaker who uttered the target word.

Finally, given a target word, processing the corpora yields a list of context words for both the Russian and the Dolgan language. Although it is expected that this list of context nouns is typically very diverse, some of the context nouns will likely occur multiple times. Precisely the words that tend to co-occur more frequently with certain target nouns, consequently, *might explain something about that target word.*

Now that a procedure for obtaining context words has been laid out, the next 'sub-subsection' will be concerned with an initial exploration into which words might make good candidates as target words.

3.1.2. A first examination of the INEL Dolgan Corpus

As touched upon in the introduction, the study is limited by practical constraints due to the data set size. Therefore, the choice of investigable target words is heavily dependent on the amount of data available in the INEL Dolgan Corpus. More specifically, the corpus must contain a target word of interest in a sufficient amount of different contexts for the resulting findings to be deemed meaningful. To give an impression of the extent of the problem, the most frequently occurring nouns are outlined in table 1. Ideally, a word of interest should be listed in this table as words occurring less than 50 times would produce too unreliable results.[8] However, it must be noted that the size of the contextual information

[8]One purpose of this thesis is to examine the options for semantic comparisons involving especially low-resource languages. Dolgan clearly is an example of a low-resource language: consider, by comparison, the Russian National Corpus, which contains more than a billion words, according to https://ruscorpora.ru/new/corpora-stat.html (Retrieved 3rd of May, 2022). Therefore, it makes sense to set the threshold considerably lower than typical for semantic analyses, *i.e.*, 300 examples (Tissari, Vanhatalo, & Siiroinen, 2019, p. 295). The threshold of 50 examples was empirically chosen following the logic that a higher threshold would simply entail that there are almost no words to be analysed.

available is considerably higher than the listed occurrence frequency, as the target word is usually embedded with *multiple* context units (*i.e.*, the target word is a single word while *all* neighboring words in the same sentence form its context); of course, this is only advantageous for data processing and visualisation, as every corpus example (typically a full sentence or a couple of sentences) could generally be regarded as representing a single context—that same context just requires multiple words to be conveyed.

Table 1 additionally illustrates the polysemic capacity of certain words in Dolgan (*e.g.* 'hir'). Without the need of further analysis, it is evident that some of these *seemingly* polysemic lemmas have a cultural explanation: 'd'ie͡' is used to refer both to a 'house' and a 'tent'; in light of the findings presented in section 2, a 'tent' and a 'house' is likely regarded to be the same thing by Dolgan people. Furthermore, the appearance of words that are descriptive of surroundings, such as 'reindeer,' 'snow', 'river', 'shore', and 'tundra', reveals the potential of language use analysis to be of anthropological value. In this light, the occurrence of *two* different words for 'tundra' in the list of most frequent words—and the apparent necessity of the existence of separate words to refer to it—is especially noteworthy (this will be addressed in more detail in §3.3). Despite their less obvious nature, words such as 'fish', 'water', 'earth', 'fire', and 'food', are also indicative of cultural traits by the mere fact that they score high among the most frequently used words in the language (likewise further discussed in §3.3).

These findings are not surprising considering the dominating weight of folklore[9] in the corpus, primarily originating[10] from the works of folklorist and ethnographer Efremov. Folkloristic texts are, inherently, culturally oriented (Danilova, 2017, p. 85); or, as Kistova et al. (2019, p. 792) puts it with regard to the same author, "of great importance in studying Dolgan worldview." Moreover, as folklore could provide a link to the past (Davidson, 1963, p. 527, 544), the folkloristic corpus texts may hold elements of the more ancient traits listed in section 2, whereas, intuitively, the non-folkloristic corpus texts stemming from recent conversations might tend to contain more references to the described, more modern way of Dolgan life; the Dolgan corpus may thus potentially grasp the entire 'range' of cultural attributes outlined in §2. Despite the large amount of folklore, the 'naturalness' of the texts of which the corpus consists can be confirmed by testing the applicability of *Zipf's law* on all distinct words in the corpus (Ellis and Hitchcock (1986, p. 426); Zipf (1949)); see figure 4. Zipf's law states that for all naturally occurring languages the frequency of occurrence of any word in a language is inversely proportional to its rank based on precisely these frequency counts. Perhaps counter-intuitively, the ubiquitous presence of folklore can be looked upon as an advantage as increasing exposure to the Russian culture

[9] Out of 76,912 words, 29,640 (≈ 38.5%) originate from folklore; the largest chunk (see footnote 7).

[10] The URL provided in the bibliography item concerning the INEL Dolgan Corpus contains a description in which a short summary is given of the contents of the corpus, in which Efremov is mentioned.

Table 1. Most frequently occurring nouns in the INEL Dolgan Corpus. In total, 68 nouns occur at least 50 times in the corpus.

Dolgan	English	occurrences
kihi	human being	1056
ogo	child	724
d’i͡e	house	418
taba	reindeer	399
d’aktar	woman	340
ogonn'or	old man	331
u͡ol	boy	293
d’ɨl	year	218
hir	place	204
ɨra:kta:gɨ	czar	202
kɨ:s	girl	194
eme:ksin	old woman	187
hir	earth	187
dogor	friend	157
kün	day	149
in’e	mother	133
is	(the) inside	133
u:	water	130
kɨ:s	daughter	129
d’i͡e	tent	128
balɨk	fish	121
a:t	(the) name	117
as	food	108
hɨrga	sled	104
u͡ol	son	99
d’on	people	96
tus	side	96
ti͡a	tundra	96
ojun	shaman	93
üle	work	93
er	husband	92
aga	father	89
uskuo͡la	school	86
mas	wood	80
taŋas	clothing	80
ürüt	upper part	79
olok	life	75
kine:s	prince	74
d’on	people	72
karak	eye	72
ilin	front part	71
ili:	hand	71
oloŋko	tale	71
te:te	father	70
kelin	back part	68
hahɨl	fox	67
er	man	64
kolxoz	kolkhoz	63
ta:s	stone	62
ehe	grandfather	62
mu͡ora	tundra	62
uraha	pole	61
ɨri͡a	song	60
aba:hɨ	evil spirit	60
gu͡orat	city	58
ičči	master part	57
ka:r	snow	57
u͡ot	fire	57
a:n	door	57
ubaj	older brother	56
at	horse	55
karčɨ	money	55
ebe	river	54
kupi͡es	merchant	54
kɨtɨl	shore	54
ma:ma	mum	53
ɨt	dog	52
alɨn	lower part	50

(or *Russification*) might have resulted in shifts in language usage, especially since contact with the Russian language has increased (Khanolainen, Nesterova, & Semenova, 2022, p. 3). The presence of words in the table such as 'czar', 'kolkhoz', 'city', and 'school' could be indicative of such Russian influences. This claim is strengthened by the fact that the former three words in Dolgan are clear loanwords from Russian (*e.g.*, Dolgan 'gu͡orat' versus Russian 'gorod' 'city'). However, although the Dolgan word for school, 'usku͡ola', shows similarities with Russian 'škola' 'school', the word is most likely etymologically linked to Turkish 'okul' 'school', as Dolgan is linked to, and possesses features of, Old Turkic (Ubrjatova, 1985, p. 17).

Whether it is evidence of a people's surroundings, remnants of past language contact, or an allusive hint of polysemy—the observation that merely a list of frequent nouns already appears to have a capacity for revealing cultural knowledge indicates the fruitfulness of linguistic corpus research as a guide for anthropology.

Thus far, this study has provided the material that will be subject to cultural semantic analysis. It is now necessary to address the method that will be used to actually conduct these semantic analyses. Taking the overview of Dolgan culture of the previous section (§2) as its point of departure, in the next section, this thesis will move on to develop such a methodology, based on semantic theory (subsection 3.2.1) and relevant existing methods (§3.2.2). Afterwards, my study will return to the question of target words and put into practice the newly acquired knowledge to finally define the experiments that will be conducted in this thesis (§3.3).

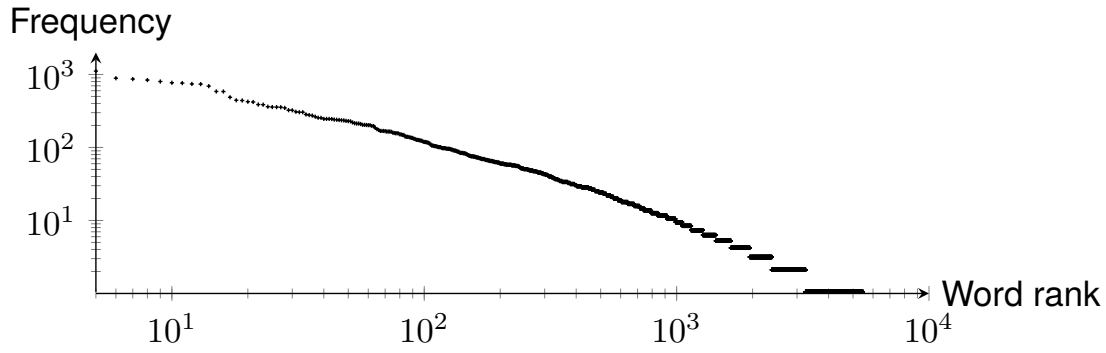


Figure 4. Log-log-scale graph of individual word frequency counts (y-axis) and frequency ranks (x-axis). Based the entire data set provided by the INEL Dolgan corpus. Due to the logarithmic axes, a language adhering to Zipf's law would produce a linear line, which is roughly the case for this graph.

3.2. Means of analysing differences in semantic meaning

An understanding of a culture can be theoretically approached through formulating it in the form of models derived from that culture (Holland and Quinn (1987, p. 4); Shore (1996, p. 44)). Such *cultural models* may arise when "we routinely, repeatedly do things with other people," such that "we develop some standardised way of doing these things" (Kecskes, 2013, p. 91).[11] In this respect, the cultural *knowledge*[12] explicitly presented in section 2 can be regarded as contained within the cultural models that exist among Dolgan people. More radically—it is solely the specific set of knowledge that Dolgans have about their world that wholly *defines* their culture (Leavitt, 2014a, p. 60).

As pointed out in the introduction, essentially, the aim of this thesis is to 'extract' such cultural knowledge from expressions of language alone; more specifically, from single *words*. The assertion of the mere possibility that this can be done already pushes the present study into a peculiar corner of linguistics. Therefore, in the first subsection, I will place this study in its theoretical context from which, in the succeeding subsection, the corresponding available methodology will naturally follow. These existing methods, and their underlying philosophies, form the 'ingredients' for the quantitative method proposed in the final subsection.

3.2.1. Theoretical framework

The first discussions of the relation between language and culture—and consequently the first comparative analyses of languages and cultures—emerged during the mid-19th century when philosopher and linguist Wilhelm von Humboldt fundamentally argued that culture and language are "so intimately fused with one another, that if one were given, the other would have to be completely derivable from it" (Von Humboldt, 1836/1999, p. 46).

Von Humboldt paved the way for Franz Boas (Leavitt, 2014b, p. 24), a pioneer of the field of anthropology, who held that language *directly reflected culture* and could therefore be used to study cultural knowledge (Lucy, 1992, pp. 13–14). Boas viewed that different languages systematise concepts that emerge from cognitive experience in different ways, based on a culturally driven notion of 'what matters' (Jakobson & Boas, 1944, pp. 190–191).[13] Following this reasoning, cultural concepts, *i.e.*, *cultural models* that emerge from

[11] Indeed it is precisely this 'characteristic, habitual, behaviour' that features in many classic definitions of 'culture' (see, *e.g.*, Keesing and Strathern (1998, p. 15) and Harris (1993, p. 104)).

[12] In the context of (cognitive) linguistics, Kecskes (2013, pp. 81–83) argues that such knowledge is mentally structured as *encyclopedic knowledge*. In this view, all knowledge that an individual has about the world is organised in the mind in the form of an interconnected network of concepts and models (the terms 'concept' and 'model' defined in the philosophical sense). This 'encyclopedia of world knowledge' is consulted *through language* such that *meaning* arises in the form of a composition of selective parts of this 'network of ideas about the world' based on the perceived linguistic sign, *i.e.*, a *word*. It is thus encyclopedic knowledge that *underlies* linguistic meaning.

a group of people's *shared* cognitive experience, are thus integrated in the structure of a culture's language. As brought forward in the introductory text to this section, it is precisely *knowledge of these cultural models* that constitutes a culture, and which is thereby, according to Boas, naturally imposed on, and conveyed by, the language spoken among individuals within the culture.

Although Boas affirmed the influence of culture on language, he was skeptical of the opposite being true (Lucy, 1992, p. 21). Inspired by Von Humboldt and Boas, (Leavitt (2014b, p. 25); Lucy (1992, p. 26)), Sapir, and especially Whorf, finally lay the foundation to complete the circle and argued for what is now widely known under the header of the *linguistic relativity principle*: the idea that thought and language are inextricably linked (Lee, 1996, p. 27).[14] The manner in which these are linked might be best explained with reference to the structuralist movement in linguistics (Lee, 1996, pp. 74–82), developed among the same 'Humboldtian stream' (Geeraerts (2010, p. 51); Leavitt (2014b, p. 23); Leavitt (2014a, pp. 53, 56–57)): structural linguistics assumes language to be a sort of 'interface tool' between concepts in the mind and experience from the world (Geeraerts, 2010, p. 51).[15] This is especially interesting when such mental concepts (or models) are considered to be *cultural concepts*, such that the underlying structure, *i.e.*, language, associating these concepts with the world, and vice-versa, comes to be organised differently—with implications for how people perceive the world.

[13] Boas' standpoint is perhaps most easily explained as follows. Different peoples make sense of their perceptual experiences in different ways: based on a shared belief of 'what is important in life,' a 'culture' emphasises certain cognitive sensations over others. Language is a sort of 'index' into this overwhelmingly large 'cloud of thoughts' and, due to economical reasons, a language system is *forced* to 'make choices.' Therefore, a language system is *shaped to accommodate for the set of priorities that a culture demands*; the set of priorities through which a culture perceives the world. The most salient characteristics of a culture (*e.g.*, an overt distinction based on gender in the form of gender roles) become embedded—as is most efficient—in a language's grammatical system (*i.e.*, via gender pronouns) such that they become obligatorily expressed, while other, likewise important, cultural characteristics become part of the language's lexicon.

[14] It is important to stress that neither Sapir (Leavitt (2010, p. 135); Lucy (1992, p. 20)), nor Whorf (Whorf, Carroll, & Chase, 1956, pp. 138–139) agreed on a direct correlation between culture and language; the argument instead pertains to the mutual relationship between language and *thought* (and thus only indirectly relates to culture through habitual behaviour guided by thought). I do also not claim that the theory holds that *all* thought is initiated by language, as is sometimes erroneously believed (Lee, 1996, p. 30).

[15] In this sense, one could roughly view that language functions as a mechanism that facilitates associations between experience (*i.e.*, the sensing of a physical object) and concepts or models in the mind (*i.e.*, a mental representation of that physical object). Also see footnote 12. It is important to note that this mechanism, language, emerges as an independent structure from which meaning arises as a product of its *whole*: a notion of 'good' is only 'meaningful' *in relation* to a notion of 'bad'; the underlying conceptual system to which language provides access needs to contain models for both in order for language to emerge with contrasting signs, *e.g.*, the words 'good' and 'bad'. In this sense, a 'culture' can be seen as a *semantic whole*, as Boas puts it, which is internally integrated into language (Leavitt, 2014a, p. 53).

This idea, that "language, thought, and culture are deeply interlocked," (Levinson & Gumpers, 1996, p. 2) did stand in stark contrast with Chomskyan ideas of *universal* linguistic principles which, inherently, deny significant diversity of language (or culture) (Leavitt, 2014b, pp. 26–27). It is therefore not surprising that rather in the legacy of Humboldt, Boas, Sapir, and Whorf, I find the point of departure for my thesis; indeed, there was not much room for *words*—the main subject of my study—in Chomskyan linguistics (Goddard & Wierzbicka, 2014, p. 5).

Actually, Humboldt, Boas, Sapir, and Whorf were all engaged in *ethnosemantics* as they dealt with meaning *across different cultures* (Leavitt, 2014a, pp. 51–53); or, alternatively, and practically synonymous, in *cognitive anthropology*, as they studied *humanity* from a cognitive perspective (Ottenheimer, 2012, pp. 22–23). After the discipline fell into neglect in the Chomsky era (Leavitt, 2014a, pp. 61–62), during which culture was often divorced from language, it was revived by Mel'uk, Goddard, Wierzbicka, and their colleagues (Goddard & Wierzbicka, 2014, p. 7), through their introduction of *neostructuralist* methods (Geeraerts (2010, pp. 124–125); see §3.2.2).

Finally, in their methods, it was *words* that came to play a central role in "everyday meaning-making" (Levisen & Waters, 2017, p. 2): "it is time to bring words back in and to recognize them as part of the core business of linguistics" (Goddard & Wierzbicka, 2014, p. 7). While cultures themselves emerge through habitual behaviour, with which this section in my thesis began, it is again *through routine and repetition* that people acquire 'a cultural feeling for language,' and thus its lexicon (Garrett, 2006, p. 605). Following Wierzbicka (2013, pp. 306–307), in her section titled *Why words matter*, precisely words contain a people's ways of thinking and shape a people's world. More specifically, as Kecskes (2013, p. 83) puts it with regard to cultural models, likewise introduced above, words are "points of access" to the (cultural) knowledge contained within.[16] In summary, "if we understand what words mean to people, we can tap into their world of knowledge, values and orientations" (Levisen & Waters, 2017, p. 4).

Especially culture-specific words, *e.g.*, Dutch *gezellig* (Peeters, 2020) or Russian *polost'* (Nabokov, 1961, pp. 64–65), bear cultural knowledge (Wierzbicka (1997); Goddard and Wierzbicka (2014, p. 8); Levisen and Waters (2017, p. 3)) in such manner that through analysis of their meaning one's worldview can be 'decoded' (Levisen & Waters, 2017, p. 8). These *cultural keywords* "can give access to the inner workings of a culture as a whole, to its fundamental beliefs, values, institutions and customs" and are thus "words that *explain* a culture" (Rigotti & Rocci, 2005, pp. 125–126). That raises the question

[16] More concretely, Kecskes builds on Langacker (1987, pp. 163–164) who argues that lexical items are the "points of access" to *encyclopedic knowledge*; see footnote 12. The findings by Wierzbicka (1985, p. 4) presented here moreover further strengthen the claim that the semantic meaning of concrete nouns indeed draws from encyclopedic knowledge.

whether the words considered in my thesis—which were purposely chosen to be everyday words that appear in the world's languages nearly universally—can validly be labeled 'culturally specific.'

Akizhanova, Zharkynbekova, and Satenova (2018, p. 80) note that cultural keywords words must occur with an unusual frequency in a corpus and, conveniently, propose to use the Zipf distribution, which I already introduced in §3.1.2 for a different purpose, to compare word frequencies across languages and consequently find potential culture-specific words. Such distributions are readily available for, *e.g.*, the Russian language (Šarov, 2001) and can thus be used to determine which Dolgan words are characteristic for Dolgan culture (at least in comparison to the Russian language).

However, as Wierzbicka (1985, pp. 4–5) crucially notices, everyday words—with which I am specifically concerned in the current study—also *differ in meaning across languages*; Wierzbicka's contributions to the contemporary field of ethnosemantics thereby lay the theoretical groundwork for the methodology used in my thesis.

It is unavoidable to conclude this subsection with a few remarks on the classic distinction between *semantics* and *pragmatics* in linguistics. As the sphere of pragmatics is generally held to concern, among other things, the *contextual* meaning of words (Leech, 1983, p. 13), one might be inclined to include the cultural dimension with which this paper engages as pragmatic context, especially since important aspects of such contextual information are a language's social and cultural surroundings (Senft, 2014, pp. 129–130). It is, however, for good reason that Bar-Hillel (1971) famously referred to the field of pragmatics as a 'wastebasket': phenomena are often labeled as 'pragmatic', while they are *actually the business of semantics* (Bach, 1997, p. 36). Perhaps similarly, the theory discussed in this part of my thesis strictly views *all* culturally shared conceptual knowledge as, core, semantic meaning (Kecskes, 2013, pp. 82–83). Furthermore, the 'boundaries' of what such semantic meaning may encompass are set by the respective cultural communities themselves (Wierzbicka, 1985, pp. 4, 214–217).[17] Semantic meaning thus differs across cultures (and hence the need for a field of ethnosemantics). Pragmatic meaning, then, ultimately pertains to contexts that vary *within* a culture (*e.g.*, speech acts), while their influence on interpretation is likewise assumed to differ across cultures and is studied in the field of ethno*pragmatics* (Goddard & Ye, 2014, p. 66). For example, while, in Dutch, 'je vader' has an equivalent translation into English: 'your father', regardless of any differences in semantics, the respective *pragmatic* interpretations of these words differ as, in English, the phrase is more often used in rhetorical settings (Beekhuizen, 2021). In summary, my thesis is concerned with semantic meaning across cultures.

[17] Among members of a culture sustained primarily by hunting, it is conceivable that conceptualisations of certain wild animals are richer; their concepts contain knowledge that refers to their cultural habits of hunting.

3.2.2. Overview of existing methods

Structuralist approaches within semantics can be broadly generalised to belong to the following three categories, each of which builds on top of another; and, all of which assume language to be an "intermediate level between the mind and the world" (Geeraerts, 2010, p. 52), as elaborated in the previous subsection (and more comprehensively in footnote 15). According to lexical field theory, conceptualisation of the world happens through the division of it into separate 'fields' that consist of sets of mutually interdependent words (Lehrer, 1974, p. 15), *e.g.*, the lexical field `FEELING` delineates the words 'sad' and 'happy', among others (Faber Benítez & Mairal Usón, 2013, p. 38). Componential analysis is then concerned with the analysis of exactly these relationships between words within the same field: they are the *semantic* features by which words within a lexical field are distinguished from one another (Ottenheimer, 2012, p. 26), *e.g.*, `NEGATIVE` versus `POSITIVE` (relevant for the previous example) or `MALE` versus `FEMALE`. Similarly, relational semantics focuses on the analysis of the *structural* relations—with which structuralism is ultimately concerned—detached from their resultant semantic meaning, such as hypernymic[18] relationships (Geeraerts, 2010, p. 52). Neostructuralist methods, first mentioned in the previous subsection, take these classical approaches as a basis and develop these further in various ways (Geeraerts, 2010, p. 124). What follows is a brief overview of the available such methodology relevant for the aim of my thesis. In §3.2.3, the techniques that will be elaborated upon in this section will be merged to form a new method, more suited for identifying semantic differences between Dolgan and Russian. In this overview, I will not expand on the details involved with actually operating the methods, but rather focus on the usefulness of their products in terms of applicability to the question addressed in this thesis.

Natural Semantic Metalanguage (NSM) is a framework that grounds itself in the belief that the meaning of all words can be decomposed into a small number of primitive factors (or *primes*) that universally occur in all natural languages and represent relationships by means of a universal syntax (Goddard & Wierzbicka, 2014, pp. 10–18)—a sort of linguistic counterpart to the fundamental theorem of arithmetic, which states that every integer number can be factored as a unique product of prime numbers (Weisstein, 1999, p. 687). In this way, the method actually combines the universalist Chomskyan perspective with the relativist Whorfian view that are normally standing in opposition to each other.

The conceptual primes include words such as `KNOW`, `THINK`, `ABOVE`, `SOMEONE`, `GOOD`, `BAD`, `BIG`, and `SMALL`, of which currently 65 are 'discovered' (Wierzbicka, 2010, pp. 7–8). A very concise example of a semantic description (technically called an *explication*)

[18] Hypernym-hyponym relationships form ontological 'taxonomies' and are best explained with an intuitive example: a quail is a *type of* bird; therefore, 'bird' is the hypernym of 'quail' and 'quail' is the hyponym of 'bird'. Note that 'bird', on its own turn, is a hyponym of 'vertebrate', which then is a hyponym of 'animal'.

is given in (1) below. Note that the word 'round' is not a primitive. However, it can by itself be expressed solely by primitive terms, and may therefore function as a *molecule* in further explications (one could intuitively view this as performing iterative reduction into primes). Proper explications encompass all of a word's meaning and are therefore much more detailed and, consequentially, lengthy (which is by itself, as Allan (2020, pp. 450–456) substantiates, also a major limitation of the method); see, *e.g.*, the explication of the English word 'cat' in Goddard (2011, pp. 206–209) or the description of the word cup in Allan (2020). However, it may be apparent that such detailed explications indeed contain cultural-specific components of the semantic meaning of certain words.

(1) *head (someone's head):*
one part of someone's body
this part is above all the other parts of the body
this part is like something round [M]
when someone thinks about something, something happens in this part of someone's body

Clearly, this method seems analogous with the aforementioned componential analysis approaches; however without strict demarcations of any lexical fields, as the only separated elements are the primes themselves, through which all of the 'fuzzy rest' can be described (Geeraerts, 2010, pp. 124, 126–127). As a direct consequence of the fundamental assumption that these primes are universal, NSMs can be constructed for *any* language, and, the 'building stones' it consists of can therefore be used to formulate universally understandable 'recipes' for the meaning of (seemingly equivalent) words *across* languages, an application that Wierzbicka (2010, pp. 8–16) herself demonstrates with a comparison between the Russian word *sud'ba* and its English translation equivalent *fate*. The NSM framework thus poses a potential method for research into (subtle) semantic differences between translation equivalent words across Dolgan and Russian. However, analysis using NSM is a labour-intensive process which, most importantly, requires at least some degree of intuition in the target language, and, moreover, the availability of a sufficient amount of linguistic data (usually originating from a corpus) to be scrutinised (Goddard, 2015, pp. 821–822); all of which were reasons for why I instead opted for a pure quantitative approach in the context of the problem posed in this thesis. The intuition to be gained here is that parsing a corpus may yield sufficient use cases (or contexts) of a word, such that the explications that may be derived from such analyses are exhaustive, and, therefore, indicative of cross-cultural differences in semantic meaning of words upon mere comparison with each other (note that differences between explications across languages reveal exactly the sought-for differences in meaning across these languages).[19]

Word embeddings are typically used within the discipline of Artificial Intelligence and, more specifically, its subfield of *Natural Language Processing*. Word embeddings are high-dimensional, numerical, representations of words that generally preserve the notion that words that are similar in meaning are closer in terms of their numerical representation (Jurafsky & Martin, 2021, pp. 96–125). Put in a slightly more formal way, if the word embeddings behave as mathematical vectors in a vector space, the distance between word embedding vectors of similar words would be smaller than between those of words that are further apart in meaning. A visualisation of such a space would look similar to figure 1—the intuition this thesis started with. However, instead of two dimensions, one should imagine several hundred dimensions in which the vectors live (note that in this figure the vectors are rather far apart). Various ways exist for obtaining word embeddings. As all of these are mathematically involved, their explanations are out of the scope of this thesis. Crucially, all of these methods rely on the *distributional semantics*-hypothesis stated in the introduction (Almeida & Xexéo, 2019, pp. 1–2); the method therefore inherently belongs to the formerly introduced relational semantics approaches (Arseniev-Koehler, 2021, pp. 12–15). Word embeddings can be obtained in many different languages (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018) and can also be used for cross-language comparisons of meaning, *e.g.*, in Luong, Pham, and Manning (2015, p. 157).

To strengthen intuition, a very basic example of word embeddings is shown in table 2. This example immediately demonstrates a major limitation of this method with regard to the aim of this thesis: as the word embeddings (`[1, 2, 2, 1, 1, 0, 0, 0, 0, 0]` and `[1, 2, 0, 0, 0, 1, 1, 1, 1, 1]` for, respectively, Dolgan and Russian) are ten-dimensional, a visualisation of the resulting vectors becomes a challenge, especially since word embeddings are much larger in dimensionality in practice. Again, many methods exist to reduce the dimensionality (*e.g.* principal component analysis (Pearson, 1901)), however, these come at the cost of interpretability—another quality that we required from the methodology—as the dimensions of the resulting embeddings will not correspond to human-interpretable concepts, while at least in table 2 the dimensions directly represent meaningful ideas relevant to the problem (namely context words). The enormous success of the method in other applications nonetheless demonstrates the potential of merely using word co-occurrence statistics to define a word's similarity in relation to others, without actually qualitatively 'capturing' the meaning, in contrast to when doing analysis within the NSM framework (Bakarov, 2018, pp. 1–2).

[19] For example, the NSM explication for 'dog' in Russian could be almost entirely in accordance with the explication for 'dog' in Dolgan, although the latter might miss a 'many animals [m] of this kind live in people's homes [m]' clause, while it could instead contain a clause hinting on the typical usage of dogs for transport. Upon comparison of these explications, the difference in meaning could simply be 'distilled'.

WordNet is a lexical database that is structured to represent a *network* of words, based on their underlying semantic relationships (Miller, 1995). WordNets have been constructed for dozens of languages[20], including Russian (see also §3.2.3). In WordNet, nouns are grouped by their synonymous relationships to form synonym sets (or 'synsets'). For example, 'dog' is part of the synset that also includes 'domestic dog' and 'Canis familiaris'. In addition, these synsets are linked to each other by hypernymic (see footnote 18), antonymic and meronymic relationships. Elaborating on the formerly introduced example: the 'dog'-synset is thus linked by a hypernym relation to a synonym set that consists of the words 'animal', 'animate being', 'beast', 'brute', 'creature', and 'fauna'. Clearly, WordNet is therefore a tool within relational semantics. However, it does not obey the structuralist paradigm in the strict sense: unlike NSM, WordNet does not claim to explicate the complete semantic meaning of words (Geeraerts, 2010, p. 160). In similar character to the previous method, semantic analyses that purely involve the structural relationships present in WordNet consequently suffer from the limitation that meaning can not be exhaustively conveyed. A visualisation of hypernymic relationships is shown in figure 5 on the next page. Intuitively, and by definition, a hypernym is a *generalisation*, while a hyponym is a *specialisation*. Looking again at figure 1, 'transport' could be a hypernym of 'sled' (in fact, it is) while 'pet' feels like a generalisation of 'dog'. Following this observation, WordNet could potentially pose a solution to the dimensionality problem of word embeddings stated

Table 2. Co-occurrence frequency based word embeddings for dog in Russian and Dolgan based on a two-sentence Dolgan corpus (A & B) and a two-sentence Russian corpus (C & D). Only nouns (marked in bold) are counted.

	human	dog	sled	ice	clothes	home	cat	cap	place	couch
dog (Dolgan)	1	2	2	1	1	0	0	0	0	0
dog (Russian)	1	2	0	0	0	1	1	1	1	1

(A) Onu iste:t kord'on kihi taksɨbɨt, ɨtɨgar olorbut da: bararga bu͡olbut.
'As he had heard this, the small **human** went out, sat down on his **dog sled** and hurried away.'

(B) "Bu:s killeste taksi͡am," d'i͡en baran tanastarɨn kɨbɨnan taha:rbɨt, ɨtɨn hɨrgatɨgar u:ran ke:spit.
"'I will go out and bring some **ice**," he said, took his **clothes** outside and threw it onto the **dog sled**.'

(C) Doma živut tol'ko koški i sobaki, -- rezonno zametil čelovek v kepočke.
'Only **cats** and **dogs** live at **home**, – reasonably assumed the **man** with the **cap**.'

(D) Ne mesto sobakam na divane valjat'sja!
'There is no **place** for **dogs** to lie on the **couch**!'

[20]Global WordNet Association. *Wordnets in the World*, http://globalwordnet.org/resources/wordnets-in-the-world/ (7th of May, 2022).

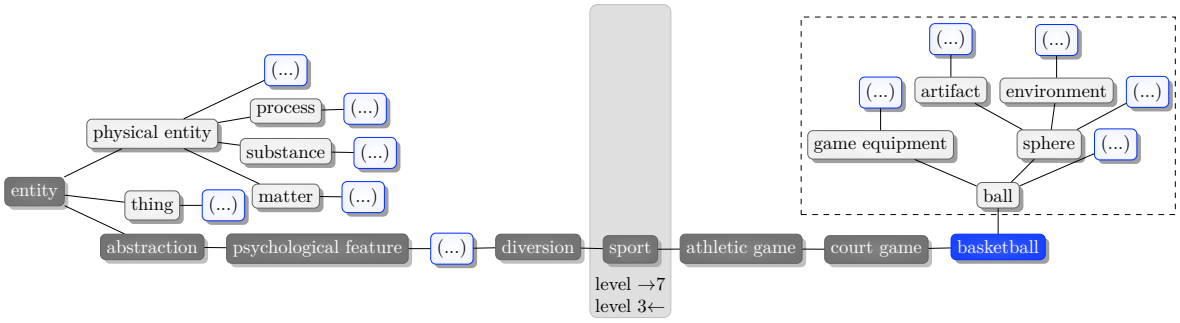


Figure 5. Hypernym tree for the word 'basketball'. The tree shown in the dashed rectangle depicts a second meaning of 'basketball': the *ball* instead of the *game*. In this subtree, even more such polysemy is encountered: *sphere* amounts to seven different sets of synonyms. All possible trees have the same root: *entity* (in fact, all hypernymic relations are represented by one, big, tree.) The grey box highlights the synset at the 7th level of the graph from the root note; note that the lower the level number, the closer to the root of the tree, and, generally, the higher the degree of 'abstraction'. Alternatively, the synset is at the 3rd level from 'basketball'. The visualised tree is truncated: not all levels are shown and expanded, and only one word is picked per synset.

previously. The fact that many words are polysemous, likewise illustrated in figure 5, is, however, a large limitation of such an approach—the problem of *word sense ambiguation* is non-trivial and remains unsolved (Navigli, 2009). For example, the word 'dog' is also part of a synset that furthermore consists of the words 'frank', 'frankfurter', 'hotdog', 'hot dog', 'wiener', 'wienerwurst', and 'weenie'.

In summary, it has been illustrated that all aforementioned methods, among which NSM provides the overarching inspiration, have insurmountable shortcomings for applying them to the problem posed in this thesis. However, *taken together*, their theories form an approach that is worth investigating: the generalising capacity of WordNet could mitigate the high dimensionality of word embeddings. The fact that the latter two approaches are unable to completely 'unravel' the semantic meaning of target words is of less relevance to the question of this thesis as their results are interpretable and meaningful for human qualitative judgement. As stated in the introduction, the purpose of the method is to merely *inform* subsequent quantitative analyses. This leads us to the introduction of a mixed method, which will be proposed in the next subsection.

3.2.3. Experimental quantitative method

Taking as the input a target word, parsing the corpora (§3.1.1) yields a list of context words as output. If counting the occurrences of each individual context word, one effectively ends up with a word embedding. However, as many context words occur only once or twice, the resulting set of words is highly varied. In other words, the number of *different words* is very large. Generalisation through hypernymy relations from a WordNet reduces the

number of different words: with every step up the tree, the context nouns become *more abstract* and *less varied*. I contacted the first author of the Russian version of WordNet, RuWordNet (Loukachevitch, Gerasimova, Dobrov, Lashevich, & Ivanov, 2016), and obtained an offline version of their database, containing 133,745 words[21]. Using a Russian WordNet eliminates the need for a further translation of words into English, which would only introduce more potential for inaccuracies.

As many words have ambiguous meanings, this might lead to multiple hypernyms; one for every sense of a word. In turn, these hypernyms might themselves be polysemous and expand into even more words. However, despite the total number of generalised words increasing, the variance within the group of context words *decreases* with every step. Would the level of abstraction be maximal, the total number of obtained words would be a multitude of the number of context words that was started out with, though with only one possible term remaining: sušč̌nost' 'entity'.

This method strongly builds on the expectation that the hypernyms of the correct sense of a context word will dominate: if three similar context words would each have two senses, one of which is not relevant, while additionally assuming that these 'wrong' senses are all very different from each other, the result would be that the 'correct' hypernym would occur three times in the resulting group, while the three 'false' hypernyms each only occur once. Therefore, the hypernyms of *all* possible synsets are collected, allowing the method to focus solely on semantic analysis, without having to take into account the problem of word sense disambiguation. However, as every step up the tree magnifies the polysemy problem, a maximum of only one step is taken (this concerns the level calculated bottom-up; which in figure 5 would mean the level derived with respect to 'basketball').

Ultimately, the group of abstracted nouns and the corresponding counts can function as an *explication*, like in NSM; although in a disparate form, and arrived at via a different, more quantitatively focused path. Following this observation, the 'explications', *i.e.*, the groups of hypernym occurrence frequencies, should match to a certain extent between Russian and Dolgan, except for the areas in which the semantic meaning between the target words differ.

Definitely, the distributions of generalised nouns will not completely correspond in the areas where they ought to match—an inevitable consequence of the challenge of working with minority languages that this thesis purposely took up. The essential idea is that significant differences in meaning, despite the 'noise' caused by a failure of the data to converge, still rise to the surface through this method. This study deliberately takes up

[21] Note that, besides nouns, this number also includes verbs and adjectives. Furthermore, each unique word can have multiple 'meanings' due to the polysemic capacity of capacity of certain words (RuWordNet includes 154,111 'meanings'. See: https://ruwordnet.ru/ru (Retrieved 1st of July, 2022).

the burden of working with scarce data; one that demands only a rough examination of the results, as a precise evaluation of statistically insignificant data would be senseless.

In line with this intuition, when 'grasping' the difference in semantic meaning of a certain word in relation to another language, the most frequently occurring hypernyms that are not among the most frequently occurring hypernyms of the other language are deemed the most relevant. For example, if the most frequent hypernym of 'dog' in both Russian and Dolgan is 'animal', this 'dimension of meaning' is considered irrelevant to the comparative semantic analysis. In my experiments, for each language, I eliminate every hypernym that occurs in the top 3 frequently occurring hypernyms of the other language, until a hypernym remains (at the top) that does not occur in the top 3 of the remaining hypernyms of the other language. This hypernym consequently constitutes a dimension of the geometric space in which the differences of meaning are visualised in section 4. As there are two languages, these visualisations are two-dimensional.

3.3. Experimental settings

The findings in §3.1.2 now gained some more substantiation. For example, the theory developed in §3.2.1 suggests that the reason for Dolgans to have multiple words for 'tundra' stems from a culturally imposed necessity to distinguish between different 'types of tundra'. In other words, the Dolgan language, through which Dolgan people conceptualise their world, provides points of access to multiple, *narrower*, conceptual models of 'tundra', whereas the English language only has one pointer to the whole cloud of mental ideas that constitute a reference frame for 'tundra'. The difference in meaning between these different words for tundra must then have emerged from a practical difference in meaning in everyday life, sufficiently significant as to have become glued into the language.

Another such example can be found further in the dictionary.[22] Apart from a general word for reindeer ('taba'; listed in table 1), and a more specific word for a reindeer calf ('tugut'), the Dolgans use completely different words to distinguish between a 'one year old reindeer' ('abɨlakaːn'), and a 'two year old reindeer' ('ikteːne'). Considering the central role of reindeer in traditional Dolgan culture (as is obvious from section 2), it is conceivable that the age of reindeer matters significantly in Dolgan daily life. The Dolgan language supposedly adapted to such crucial differences by the emergence of different, separate nouns. A similar argument can be made for the Dolgan word that refers specifically to an 'infertile female reindeer' ('maːŋkaːj'): the infertility of a (female) reindeer likely had a significant impact on the community. Other words include, among at least a

[22] The INEL Dolgan Corpus' dictionary can be found at https://inel.corpora.uni-hamburg.de/DolganCorpus/dictionary/dolgan (Retrieved 29th of June, 2022).

god, **čelovek**, vremja, vopros, **žizn'**, delo, raz, Rossija, **rabota**, **den'**, **rebenok**, strana, mir, sistema, **mesto**, **ruka**, slovo, **dom**, slučaj, **škola**, **storona**, **ženščina**, problema, otnošenie, Moskva, lico, vid, obraz, fil'm, **den'gi**, **gorod**, **čast'**, **ditja**, **glaz**, rezul'tat, uroven', razvitie, rešenie, sila, zadača, konec, vozmožnost', golova, jazyk, oblast', vlast', process, forma, číslo, situacija

Figure 6. Top 50 most frequent nouns of the Russian language in descending order, based on the 'one million tokens'-version of the Russian National Corpus. All nouns that overlap with the 68 most frequently occurring nouns in the INEL Dolgan corpus are marked in bold. All nouns that overlap with the 100 most frequently occurring nouns in COCA are underlined.

dozen more, a 'free running domestic reindeer' ('delemiče:'), a 'tamed reindeer' ('a:ku'), and a 'reindeer bull with bare antlers' ('n'eŋče:n').

Section 3.2.1 additionally provides a tool for verifying the 'oddity' of certain words, such as 'fish' and 'fire', being found in Dolgan texts and transcripts with high frequency. In order to ground the claim that the observed frequencies in the Dolgan corpus are indeed unusual, I have generated a similar list of most frequently occurring nouns for the Russian language (using the Russian National Corpus; see §3.1.1. The list with the top 50 most frequent nouns is shown in figure 6 below. It is immediately apparent that the list is similar to the one provided by Šarov (2001). The differences that exist between the Russian frequency lists are likely the result of corpus dynamics. The reason for the Russian words that do not overlap with the Dolgan list in table 1 might, indeed, be partly attributable to cultural influences (*e.g.*, 'rabota' 'work', 'fil'm' 'movie', and, of course, 'Rossija' 'Russia', and 'Moskva' 'Moscow'). Vice-versa, the 'oddity' of the Dolgan words from table 1 that are not listed in figure 6 might be indicative of them being interesting target words to analyse.

Similar to the English language[23], words connected with time feature abundantly in the top part of the list. Moreover, most of the Russian nouns also occur in the top 100 most frequent nouns in English, based on a list[24] derived from the Corpus of Contemporary American English (COCA) (Davies, 2010), while nearly all of them feature in the top 250 (except for 'jazyk' 'language', and 'oblast'' 'region', besides 'Rossija' and 'Moskva'). These findings might suggest that the relatively low degree of overlap with Dolgan nouns is caused by cultural differences between sedentary and nomadic peoples that are conveyed by their languages.

Finally, based on the cultural description (given in §2), the contents of the corpus (§3.1.2), and the findings of this section, the words that will be investigated using the

[23]CBS News. *Study: 'Time' Is Most Often Used Noun*, https://www.cbsnews.com/news/study-time-is-most-often-used-noun/ (5th of May, 2022).

[24]Retrieved from https://www.wordfrequency.info/samples/wordFrequency.xlsx (7th of May, 2022).

proposed quantitative method can be defined. The most neutral approach would be to investigate *all* nouns that appear in table 1 (or selecting words to investigate in anywise other than by manually picking target words). However, the exploratory nature of this thesis calls for a focus on a *few interesting words* that are roughly inspected, instead of a more thorough examination that aims to reach conclusive findings on the Dolgans and their language.

The case study conducted in this thesis will put special focus on the semantic differences between Russian and Dolgan of the following words: (1) 'd'i͡e' 'dom' 'house', since it poses an interesting case as it concerns a highly frequent noun in both languages; (2) 'hahɨl' 'lisa' 'fox', as it is a frequently occurring cultural-specific word for the Dolgan language, based on the method by Akizhanova et al. (2018).

4. Results

The previous section (§3.3) has provided analyses solely on the basis of the corpus contents. In this section, using the method described in §3.2.3, I will delve deeper into the semantic meaning of the two aforeselected words: 'house', and 'fox'. The present results are significant in two major respects: it confirms that a larger window size leads to less relevant results; and, more importantly, it shows that the method succeeds in extracting cultural values from the contexts of the investigated words.

For most English translations of the dimensions, the English WordNet was used, with which many synsets within the Russian WordNet were linked.

House In Dolgan, the most significant cultural dimension is 'šest (palka)' 'pole'. In Russian, it is 'žiloe pomeščenie' 'living accommodation'. Figure 7 indicates that for Russian speakers, a house is mostly associated with a place for living. Dolgans also associate it with a living accommodation, although more commonly with 'poles'. A straightforward explanation: Dolgan people traditionally lived in nomadic tents; and, poles are used to support such tents. A note of caution is due here since 'd'ie͡' 'house' is synonymous for 'tent'. However, considering the cultural background, in this instance, the two senses of the word only strengthens the claim that for ancient Dolgans, a tent and a house were the same thing. Interestingly, if increasing the window size from 5 to 10, the most significant dimension for the Russian language becomes 'promežutok vremeni' 'time interval'. This might indicate that a house is viewed as a place in which to spend time. The finding furthermore confirms that a larger window size lowers the significance of the results, as 'time interval' clearly is less telling than 'living accommodation'. Letting the window uncapped by sentence or speaker boundaries did not yield any significantly different results.

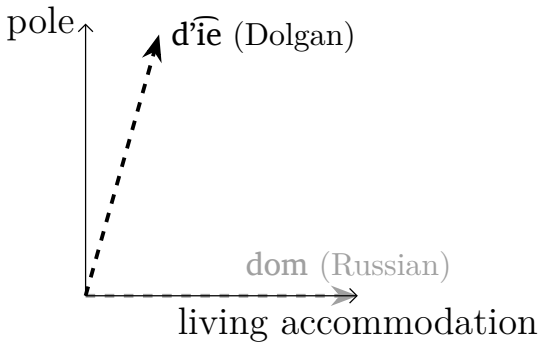
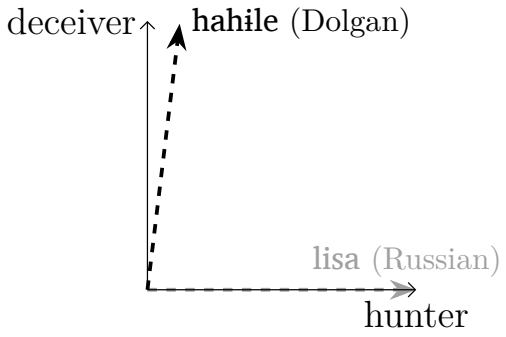
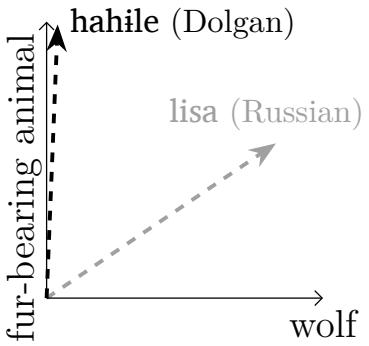


Figure 7. Context space for 'house'. In Dolgan, 5.64% of all context words were generalised into 'pole', while 1.57% of all context words abstracted into 'living accommodation'. In Russian, the words respectively amounted to 0% and 1.32%. All vectors are based on the relative frequency of occurrence within the corresponding language (% of total distribution of hypernyms) and scaled to unit size. To collect the context words, a window of size 5 was used that was further limited by sentence and speaker boundaries. For the Dolgan corpus, 216 context words were collected around 119 occurrences of 'house'. In Russian, 1412 context words and 837 instances.

Fox In Dolgan, the most significant cultural dimension is ‘obmanščik’ ‘deceiver’. In Russian, it is ‘oxotnik' ‘hunter’. Figure 8 (on the next page) indicates that for Russian speakers, a fox is mostly associated with hunting; while, for Dolgans, it is more commonly associated with a deceiving character (although also with hunting). These results are likely to be related to Dolgan folklore, in which a fox is often depicted as a ‘cheater’ or a ‘trickster’ (see section 2.4). It is furthermore conceivable that Russians mainly associate foxes in the context of hunting them (with the aim of gathering fur, or for purposes of leisure). While Dolgans are primarily sustained by hunting activities (which explains the few ‘hunter’-related contexts that were present), they originally hunted for different prey. As the Russian corpus yielded only 7 instances of the word, these results need to be interpreted with caution; it is likely that a larger corpus would have yielded at least a few Russian ‘deceiver’-like contexts. Surprisingly, when increasing the window size to 10, the most distinguishing dimension for Dolgan becomes ‘nesoveršennoletnie deti’ ‘minor children’, for which I can not offer an explanation. When subsequently not letting sentence boundaries limit the sliding window, interestingly, a whole new result is obtained (see figure 8b). Now, the most distinguishing dimension for Russian is ‘volk (životnoe)’ ‘wolf’, while for Dolgan it becomes ‘pušnoj zver'’ ‘fur-bearing animal’. A possible explanation for this might be found in the results created with the smaller window size (figure 8a): as Russians mostly use the word ‘fox’ in hunting-related contexts, it is convincing that wolves also feature in conversations or texts about hunting, although ‘a few more words away’ from foxes. Again, this shows that increasing the size and freedom of the sliding window leads to less directly relevant results. Finally, it seems possible that the emergence of ‘fur-bearing animals’ as a relevant dimension for Dolgan people is related to the fact that Dolgan people also hunt for fur-bearing animals; obviously, a fox is a fur-bearing animal. Perhaps the most striking aspect of this context space is that Russians likewise often associate foxes with ‘fur-bearing animals’—after all, it were the Russians who convinced the Dolgans to hunt for fur (see section 2.1.



(a) Window of size 5, capped sentences. In Dolgan, 4.57% of all context words were generalised into 'deceiver', while 0.57% of all context words abstracted into 'hunter'. In Russian, the words respectively amounted to 0% and 6.25%. For the Dolgan corpus, 45 context words were collected around 36 occurrences of 'fox'. In Russian, 13 context words and 7 instances.



(b) Window of size 10, uncapped sentences. In Dolgan, 2.94% of all context words were generalised into 'deceiver', while 0.11% of all context words abstracted into 'hunter'. In Russian, the words respectively amounted to 1.43% and 2.14%. For the Dolgan corpus, 215 context words were collected around 36 occurrences of 'fox'. In Russian, 47 context words and 7 instances.

Figure 8. Context spaces for 'fox', using different parameter settings. All vectors are based on the relative frequency of occurrence within the corresponding language (% of total distribution of hypernyms) and scaled to unit size.

5. Conclusion and discussion

This work has taken an exploratory step into a quantitative direction for drawing anthropological insights on the basis of language. Through experimenting with a novel method, this thesis has shown that culturally induced differences in semantic meaning between translation equivalent words in Dolgan and Russian can be identified despite being severely limited by the amount of linguistic data available.

Being limited by this scarcity of data, this study was forced to settle on a few 'band-aid solutions'. The requirement that target nouns had to occur more than fifty times in the Dolgan corpus was more or less arbitrarily set. Similarly, the decision to discover the most distinguishing hypernym by maintaining a 'top list' of three words was likewise based on empirical evidence. Also related to the low amount of available data is the fact that a Dolgan WordNet was not available. However, while it is imaginable that using a Russian WordNet for a Dolgan corpus poses a bias, it actually allows for a more fair comparison: differences in the resulting distributions of hypernyms become purely indicative of differences in contextual usage, since the same WordNet is used for both corpora. In broader terms, the major limitation of the method presented in this study is in its foundation: practically, the method is made up by merging several—very different—existing methods. Every additional procedure brings about its own limitations. For example, if a certain noun does not have a corresponding entry in the Russian WordNet, the noun is not taken into account in the final results. Many other such limitations have been discussed in section 3.2.2.

Nevertheless, the biggest contribution of this thesis is a reaffirmation of the knowledge that language bears cultural values, and, thereby, a re-emphasis on the need for preserving these languages, and consequently the cultures, that are all too often endangered. More research effort put into the preservation of such languages could naturally lead to the availability of more data, and, consequently, better methods.

In specific light of the aforementioned constraints imposed by small indigenous data sets, this paper can additionally be viewed as an overview of existing literature as it explored the possibilities for mitigating the 'small data problem' and therewith also provided a discussion on the capacity of existing quantitative methodology to be utilised for research into the semantics of low-resource languages.

Furthermore, in accordance with the elementary nature of this thesis, a case study approach was adopted, aimed at a purely exploratory analysis. In keeping with this objective, further sacrifices inevitably had to be made. To maintain exclusive focus on the case study for testing the semantic analysis method in a theoretical setting, an underlying problem of word sense ambiguation was 'tossed aside' in my research. A more sophisticated study would additionally focus on mitigating the negative effects on the results due

to the presence of polysemous words. And, although the given anthropological context of Dolgan people allowed for the experiments to be meaningful in the first place, the fact that the cultural knowledge was *a priori* given simultaneously poses a bias: the 'answers' were essentially 'given in advance', leading to a more targeted search through the resulting data. In the ideal case, anthropological accounts would be drawn *from these results*, only *after* which it should be tested by existing accounts, if available.

Moreover, although the cultural overview given in this paper was evidently, indirectly, established by native Dolgan people, studies like the present one could always benefit from access to a native Dolgan speaker. As I miss intuition in the Dolgan language, I could have very well missed certain important points in my analyses. A good example of pieces in this work that could have been aided by Dolgan speakers are the sections that were concerned with interpreting the corpus content in light of its cultural context (*i.e.*, sections 3.1.2 and 3.3).

There are a number of other remarks relating to the present study. Firstly, the Russian language is by itself spoken by many peoples from different cultures. Due to various historical reasons, it is therefore unjustified to speak about a homogeneous 'Russian culture' when referring to the Russian language. For this study, this problem is less of an issue as it mainly focuses on Dolgan culture through Dolgan language. Secondly, as the Dolgans sprang from the Yakuts (among other ancestors), their folklore is not entirely representative of 'post-Yakut' Dolgan culture; *i.e.*, some elements in the folkloristic texts might be remnants of Yakut culture and have never been a part of Dolgan life. Thirdly, the texts of the Russian corpus consist mainly of *written* language, while the Dolgan corpus almost entirely consists of *spoken* language. Many differences exist between written and spoken language, the details of which are out of the scope of this research. As such, the discrepancy between language type could have had an impact on the presented results. However, generally, as solely nouns are picked as 'key words', the potential effects of this problem are likely small, in contrast to when more complex (grammatical) structures would have been taken into account. Lastly, translation errors might be present in both the Russian and Dolgan corpora, which could have a slight impact on the final results. However, as all context words are eventually generalised using hypernymy relationships, ultimately, correct translations of words likely outweigh the incorrectly translated ones, which already reduces the negative effects due to the problem.

As a final note, essentially, the question posed in this thesis follows from the problem that words are ultimately untranslatable, despite having been marked as translation equivalents. Yet, this study's attempt to come closer to the actual meaning of these (target) words itself relies on translating (context) words (into Russian). How can translations of context words—by themselves inaccurate as per this theory—help mitigate the untranslatability of *another*, untranslatable word? This seemingly paradoxical phenomenon lies at the heart of

structuralism, from which today's successful applications of the distributional hypothesis also draw its power: how can a word be defined by its context, if all contexts by themselves are undefined? Following the theory of this paper, language constitutes a coherent system in which words only acquire meaning *in relation to other words*. Regarding the methodology discussed in this paper, it is perhaps the assemblage of context words, each of which contributes a piece to the puzzle, through which the deviant aspects of meaning can be distilled.

6. References

Akizhanova, D., Zharkynbekova, S., & Satenova, S. (2018). The zipfs law and other ways of identifying culture-specific linguistics units. *Space and Culture, India*, *6*(2), 78–93.

Alekseevič, A. N., & Efremov, P. E. (2008). Fol'klor dolgan. In *tnografija i fol'klor narodov sibiri* (pp. 427–439).

Allan, K. (2020). On the semantics of cup. In H. Bromhead & Z. Ye (Eds.), *Meaning, life and culture: In conversation with Anna Wierzbicka* (1st ed., pp. 441–460). Canberra: ANU Press.

Almeida, F., & Xexéo, G. (2019). *Word embeddings: A survey.* arXiv.

Arkhipov, A. (2020). Inel corpora general transcription and annotation principles. *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology*.

Arseniev-Koehler, A. (2021). *Theoretical foundations and limits of word embeddings: what types of meaning can they capture?* arXiv. (Preprint, 2107.10413)

Bach, K. (1997). The semantics-pragmatics distinction: What it is and why it matters. In E. Rolf (Ed.), *Pragmatik: Implikaturen und sprechakte* (pp. 33–50). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bakarov, A. (2018). A survey of word embeddings evaluation methods. (Preprint, 1801.09536)

Bar-Hillel, Y. (1971). Out of the pragmatic wastebasket. *Linguistic Inquiry*, *2*(3), 401–407.

Beekhuizen, B. (2021). Not your dad, maar wel je vader. In N. van der Sijs, L. Fonteyn, & M. van der Meulen (Eds.), *Wat gebeurt er in het Nederlands?!* (pp. 285–289). Gorredijk: Sterck De Vreese.

Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Michigan: University of Michigan Press.

Bettu, L. D. (2011). Tradicionnye zaprety dolgan. *Vestnik RGGU. Serija: Literaturovedenie. Jazykoznanie. Kul'turologija*, *71*(9), 287–297.

Bugaeva, K. M. (2013). Unikal'nyj narod dolgany. *Artika i Sever*(12), 78–84.

Danilova, N. K. (2017). pičeskij mir i" sakral'naja topografija" u dolgan. *Filologija i čelovek*(1), 85–96.

Darkgamma. (2014). *What do all languages have in common? [online forum comment].* Retrieved from https://linguistics.stackexchange.com/a/8565 (26th of June, 2022)

Davidson, E. H. R. (1963). Folklore and man's past. *Folklore*, *74*(4), 527–544.

Davies, M. (2010). The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, *25*(4), 447–464.

Davydov, V. N. (2016). Dolgany vostočnogo tajmyra: opyt polevyx issledovanij v poselkax novorybnoe i syndassko v 2015 g. *Materialy polevyx issledovanij MA RAN*(16), 67–80.

Diachkova, G. (2001). Indigenous peoples of russia and political history. *Canadian Journal of Native Studies*, *21*(2), 217–233.

Dixon, R. (2014). Basics of a language. In N. J. Enfield, P. Kockelman, & J. Sidnell (Eds.), *The cambridge handbook of linguistic anthropology* (p. 2947). Cambridge: Cambridge University Press.

Däbritz, C. L., Kudryakova, N., & Stapert, E. (2019). Inel dolgan corpus. In B. Wagner-Nagy, A. Arkhipov, A. Ferger, D. Jettka, & T. Lehmberg (Eds.), *The inel corpora of indigenous northern eurasian languages.* Hamburg: Hamburger Zentrum für Sprachkorpora. Retrieved from http://hdl.handle.net/11022/0000-0007-CAE7-1 (July 1st, 2022)

Efremov, P. E. (2000). Fol'klor dolgan. *Pamjatniki fol'klora narodov Sibiri i Dal'nego Vostoka*, *19*.

Ellis, S. R., & Hitchcock, R. J. (1986). The emergence of zipf's law: Spontaneous encoding optimization by users of a command language. *IEEE Transactions on Systems, Man, and Cybernetics*, *16*(3), 423-427.

Faber Benítez, P., & Mairal Usón, R. (2013). The paradigmatic and syntagmatic structure of the lexical field of feeling. *Cuadernos de Investigación Filológica*, *23*, 35–60.

Furniss, E. (2013). Using a corpus-based approach to Russian as a foreign language materials development. *Russian Language Journal*, *63*, 195–212.

Garrett, P. (2006). Language socialization. In K. Brown (Ed.), *Encyclopedia of language linguistics* (2nd ed., pp. 604–613). Oxford: Elsevier.

Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford: Oxford University Press.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj II (Ed.), *Language development* (pp. 301–334). New Jersey: Lawrence Eribaum.

Goddard, C. (2011). *Semantic analysis: A practical introduction*. Oxford: Oxford University Press.

Goddard, C. (2015). The natural semantic metalanguage approach. In *The oxford handbook of linguistic analysis* (2nd ed., pp. 817–841). Oxford: Oxford University Press.

Goddard, C., & Wierzbicka, A. (2014). *Words and meanings: lexical semantics across domains, languages, and cultures* (1st ed.). Oxford: Oxford University Press.

Goddard, C., & Ye, Z. (2014). Ethnopragmatics. In F. Sharifian (Ed.), *The routledge handbook of language and culture* (1st ed., p. 66-83). London: Routledge.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018, May). Learning word vectors for 157 languages. In *Proceedings of the eleventh international conference*

on language resources and evaluation (LREC 2018). Miyazaki: European Language Resources Association (ELRA).

Gries, S. T. (2013). *Statistics for linguistics with r: A practical introduction*. Berlin: De Gruyter Mouton.

Harris, M. (1993). *Culture, people, nature: An introduction to general anthropology* (6th ed.). New York: HarperCollins College Publishers.

Holland, D., & Quinn, N. (1987). *Cultural models in language and thought*. Cambridge: Cambridge University Press.

Jakobson, R., & Boas, F. (1944). Franz boas' approach to language. *International Journal of American Linguistics*, *10*(4), 188–195.

Joseph, B. D. (2008). The editor's department: Last scene of all... *Language*, *84*(4), 686–690.

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd [Draft of Dec 29, 2021] ed.). Retrieved from https://web.stanford.edu/~jurafsky/slp3/ (1st of July, 2022)

Kecskes, I. (2013). Encyclopedic knowledge, cultural models, and interculturality. In *Intercultural pragmatics* (pp. 81–104). Oxford: Oxford University Press.

Keesing, R., & Strathern, A. (1998). *Cultural anthropology: A contemporary perspective* (3rd ed.). Orlando: Harcourt Brace College Publishers.

Khanolainen, D., Nesterova, Y., & Semenova, E. (2022). Indigenous education in russia: opportunities for healing and revival of the mari and karelian indigenous groups? *Compare: A Journal of Comparative and International Education*, *52*(5), 768–785.

Kistova, A. V., Pimenova, N. N., Reznikova, K. V., Sitnikova, A. A., Kolesnik, M. A., & Xudonogova, A. E. (2019). Religion of dolgans, nganasans, nenets and enets. *Journal of Siberian Federal University. Humanities Social Sciences*, *12*(5), 791–811.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford: Stanford university press.

Leavitt, J. (2010). *Linguistic relativities: Language diversity and modern thought.* Cambridge: Cambridge University Press.

Leavitt, J. (2014a). Ethnosemantics. In F. Sharifian (Ed.), *The routledge handbook of language and culture* (1st ed., pp. 51–65). London: Routledge.

Leavitt, J. (2014b). Linguistic relativity: Precursors and transformations. In F. Sharifian (Ed.), *The routledge handbook of language and culture* (1st ed., pp. 18–30). London: Routledge.

Lee, P. (1996). *The whorf theory complex: A critical reconstruction.* (No. 81). Amsterdam: John Benjamins Publishing Company.

Leech, G. N. (1983). *Principles of pragmatics*. London: Longman.

Lehrer, A. (1974). *Semantic fields and lexical structure* (No. 11). Amsterdam: North-Holland Publishing Co.

Levinson, S. C., & Gumpers, J. J. (1996). Introduction: linguistic relativity re-examined. In S. C. Levinson & J. J. Gumpers (Eds.), *Rethinking linguistic relativity* (pp. 1–18). Cambridge: Cambridge University Press.

Levisen, C., & Waters, S. (2017). How words do things with people. In *Cultural keywords in discourse.* (pp. 1–19). John Benjamins Publishing Company.

Lindquist, H. (2009). *Corpus linguistics and the description of english*. Edinburgh: Edinburgh University Press.

Loukachevitch, N. V., Gerasimova, A. A., Dobrov, B. V., Lashevich, G., & Ivanov, V. V. (2016). Creating Russian wordnet by conversion. In *Computational linguistics and intellectual technologies* (pp. 405–415).

Lucy, J. A. (1992). *Language diversity and thought: A reformulation of the linguistic relativity hypothesis*. Cambridge: Cambridge University Press.

Luong, T., Pham, H., & Manning, C. D. (2015, June). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st workshop on vector space modeling for natural language processing* (pp. 151–159). Denver: Association for Computational Linguistics.

Miller, G. A. (1995, nov). Wordnet: A lexical database for english. *Commun. ACM*, *38*(11), 3941.

Momzikova, M. (2013). "severnaja maslenica": Den' olenevoda v sovetskom i v sovremennom rossijskom kontekstax. In A. S. Arxipova (Ed.), *Mifologičeskie modeli i ritual'noe povedenie v sovetskom i postsovetskom prostranstve* (pp. 227–237). Moscow: Rossijskij gosudarstvennyj gumanitarnyj universitet.

Montes, M. (2021). *Cloudspotting: visual analytics for distributional semantics* (Unpublished doctoral dissertation). University of Leuven, Leuven.

Morris, C. (1938). *Foundations of the theory of signs* (Vol. 1) (No. 2). Chigago: The University of Chicago Press.

Moseley, C., & Nicolas, A. (2010). *Atlas of the world's languages in danger*. Paris: Unesco.

Nabokov, V. (1961). *Nikolai gogol*. New York: New Directions.

Navigli, R. (2009, feb). Word sense disambiguation: A survey. *ACM Comput. Surv.*, *41*(2).

Ottenheimer, H. (2012). *The anthropology of language: An introduction to linguistic anthropology* (3rd ed.). Boston: Cengage Learning.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572.

Peeters, B. (2020). Gezellig: A dutch cultural keyword unpacked. In H. Bromhead & Z. Ye (Eds.), *Meaning, life and culture: In conversation with Anna Wierzbicka* (1st ed., pp. 61–84). Canberra: ANU Press.

Popov, A. A. (1934). Materialy po rodovomu stroju dolgan. *Sovetskaja etnografija*, *6*, 116–139.

Popov, A. A. (1981). Šamanstvo u dolgan. *Problemy istorii obščestvennogo soznanija aborigenov Sibiri*, 253–264.

Rigotti, E., & Rocci, A. (2005). From argument analysis to cultural keywords (and back again). In F. H. van Eemeren & P. Houtlosser (Eds.), *Argumentation in practice* (Vol. 2, pp. 125–142). Amsterdam: John Benjamins Publishing.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, *20*, 33–53.

Senft, G. (2014). *Understanding pragmatics*. Oxfordshire: Taylor & Francis.

Shore, B. (1996). *Culture in mind: Cognition, culture, and the problem of meaning*. Oxford: Oxford University Press.

Siegl, F., & Rießler, M. (2015). Uneven steps to literacy. In H. F. Marten, M. Rießler, J. Saarikivi, & R. Toivanen (Eds.), *Cultural and linguistic minorities in the russian federation and the european union: Comparative studies on equality and diversity* (pp. 189–230). Cham: Springer International Publishing.

Stolbovoi, V., & McCallum, I. (2002). *Land resources of Russia.* Laxenburg, Austria: International Institute for Applied Systems Analysis and the Russian Academy of Science. Retrieved from https://webarchive.iiasa.ac.at/Research/FOR/{R}ussia_cd/index.htm (May 9th, 2022)

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks: SAGE Publications.

Tissari, H., Vanhatalo, U., & Siiroinen, M. (2019). From corpus-assisted to corpus-driven nsm explications: The case of finnish viha (anger, hate). *Lege artis*, *4*(1), 290–334.

Ubrjatova, E. I. (1985). Jazyk noril'skix dolgan.

Von Humboldt, W. (1836/1999). *On language: On the diversity of human language construction and its influence on the mental development of the human species*. Cambridge: Cambridge University Press.

Šarov, S. A. (2001). *Častotnyj slovar'*. Rossijskij NII iskusstvennogo intellekta. Retrieved from http://www.artint.ru/projects/frqlist.php (April 3rd, 2022)

Weisstein, E. W. (1999). *Crc concise encyclopedia of mathematics* (1st ed.). Boca Raton: CRC Press.

Whorf, B. L., Carroll, J. B., & Chase, S. (1956). *Language, thought, and reality: Selected writings of benjamin lee whorf* (13th ed.). Cambridge: Technology Press of Massachusetts Institute of Technology.

Wierzbicka, A. (1985). *Lexicography and conceptual analysis*. Ann Arbor: Karoma.

Wierzbicka, A. (1997). *Understanding cultures through their key words: English, russian, polish, german, and japanese*. Oxford: Oxford University Press.

Wierzbicka, A. (2010). Cross-cultural communication and miscommunication: The role of cultural keywords. *Intercultural pragmatics*, *7*(1), 1–23.

Wierzbicka, A. (2013). Kinship and social cognition in australian languages: Kayardild and pitjantjatjara. *Australian Journal of Linguistics*, *33*(3), 302–321.

Zakharov, V. (2013). Corpora of the Russian language. In I. Habernal & V. Matoušek (Eds.), *Text, speech, and dialogue* (pp. 1–13). Berlin, Heidelberg: Springer Berlin Heidelberg.

Zamarayeva, Y. S., Kistova, A. V., Pimenova, N. N., Reznikova, K. V., & Seredkina, N. N. (2015). Taymyr reindeer herding as a branch of the economy and a fundamental social identification practice for indigenous peoples of the siberian arctic. *Mediterranean Journal of Social Sciences*, *6*(3 S5).

Zipf, G. K. (1949). Human behavior and the principle of least effort: an introd. to human ecology.