Non-canonical SAY in Siberia: Areal and genealogical patterns

Dejan Matić, Brigitte Pakendorf

Non-canonical SAY in Siberia: Areal and genealogical patterns

[article published as: Matić, Dejan & Brigitte Pakendorf (2013): Non-canonical SAY in Siberia: Areal and genealogical patterns. *Studies in Language* 37 (2): 356-412; https://doi.org/10.1075/sl.37.2.04mat]

Dejan Matić
Max Planck Institute for Psycholinguistics, Nijmegen
dejan.matic@mpi.nl

Brigitte Pakendorf
Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon Lumière 2
brigitte.pakendorf@cnrs.fr

Non-canonical SAY in Siberia: Areal and genealogical patterns*

Abstract

The use of generic verbs of speech in functions not related to their primary meaning, such as to introduce complements or adjuncts, is cross-linguistically widespread; it is also characteristic of some languages of Siberia. However, the distribution of non-canonical functions of generic verbs of speech among the languages of Siberia is very heterogenous, with striking differences even between dialects of one language. In this paper we attempt to elucidate whether shared inheritance, parallel independent developments, or areal convergence are the factors determining this distribution, using fine-scaled investigations of narrative data from a large number of Siberian languages and dialects. This enables us to uncover a wide range of non-canonical functions that the generic verb of speech has acquired in some of the languages investigated, as well as to highlight the very complex historical processes at play.

1. Introduction

Anderson (2004, 2006) has suggested that the languages of Siberia might constitute a linguistic area defined by a number of shared phonological and morphological criteria. A further feature that many languages of this vast territory share is the use of generic

* The analyses for this paper were initiated while we were both still affiliated with the Max Planck Research Group on Comparative Population Linguistics at the MPI for Evolutionary Anthropology, Leipzig, Germany. The Dolgan and the Ėven data were collected under the auspices of this research group with funding by the Max Planck Society; the collection of Ėven data was continued in the framework of a DoBeS project on the cultural and dialectal diversity of Ėven with funding by the Volkswagen Foundation. The Sakha data were collected with financial assistance of the Wenner-Gren Foundation for Anthropological Research, Inc. and the MPI for Evolutionary Anthropology. A very preliminary version of this paper was presented at the Department of Linguistics, MPI for Evolutionary Anthropology, in September 2009; we thank our colleagues as well as three anonymous reviewers and Balthasar Bickel for their helpful comments and feedback, which greatly improved our anlysis. We are extremely grateful to the linguists who shared their unpublished narrative corpora with us: Irina Nevskaya, Sofia Oskolskaya, Eugénie Stapert and Andrej Shluinsky; this study would have been much the poorer without their generosity. Thanks are also due to Michael Dunn, Olesya Khanina, Gerson Klumpp, Larisa Leisiö, Irina Nikolaeva and Elena Skribnik, who answered our questions about SAY in various languages and gave us access to their unpublished or inaccessible papers. Remaining errors are solely our responsiblity.

verbs of speech (forthwith called SAY) in a variety of functions not related to speech acts, such as the marking of a purpose clause illustrated in (1).

(1) Kolyma Yukaghir (Nikolaeva 2004: 38.12)

Čoyojə-pul (...)	*norqəyə-nu-ŋi*	*əl=šejr-ej-gə-n*	***mon-u-t***
knife-PL	jerk-IPF-3PL	NEG=escape-PF-IMP-3SG	say-0-IPF.CVB

'The knives were moving (...) in order to prevent me from going out.'

However, expanded uses of SAY (henceforth: non-canonical SAY) are cross-linguistically very common[1], and many of the languages of Siberia that share them belong to the Turkic and Mongolic language families, in which non-canonical SAY can be argued to be inherited (cf. Section 2.3). Hence, the obvious question is whether there is any reason to assume an areal nature of the phenomenon illustrated in (1): if non-canonical SAY is cross-linguistically common, can it not be the case that it arose in a number of Siberian languages independently, since the path from SAY to a purposive conjunction, for example, is a 'natural' one? And if many languages spoken in this territory have inherited this feature, why not assume that the frequency of SAY is just a historical accident, in which the speakers of non-canonical-SAY languages happened to spread over a certain area?

Furthermore, our data show that the seemingly clear-cut areal picture is less monolithic when enough details are taken into account: non-canonical SAY is well attested in central and southern Siberia, but is absent in the west and the extreme northeast. Particularly striking are the notable differences among dialects of individual languages, as illustrated in Table 1 for the North Tungusic languages Ėven and Evenki:

[1] Non-canonical SAY occurs in Africa (cf. Güldemann 2008 and the references therein), Southeast, East, and South Asia (e.g. Chappell 2008, Genetti 2011), Melanesia (e.g. Reesink 1993, Klamer 2000, Aikhenvald 2009: 388-389), and the New World (e.g. Adelaar 1990, Broadwell 1991: 421ff).

Table 1: Frequency of non-canonical SAY in dialects of Ėven and Evenki

	Corpus size	Tokens of SAY	Non-canonical SAY
Western Ėven	~54,800 words	1081	32.5% (351)
Eastern Ėven	~51,700 words	608	0.3% (2[2])
Western Evenki	~16,100 words	476	0% (0)
Eastern Evenki	~11,800 words	529	12,5% (66)

As we will show here with an analysis of narrative corpora (cf. Section 2.1), the Siberian SAY-'area' turns out to be a set of overlapping micro-areas, and the simple dichotomy *inheritance vs. contact* must be replaced by a more complex network of mutal feature exchanges, unequally distributed inheritance of features and, occasionally, independent parallel developments. The work with corpora enables us to find evidence for linguistic structures which are rarely described in grammars[3]; in addition, using these text data we are able to rely not only on the traditional method of formal comparison, but to also make use of frequency counts for exploratory data analysis that lets us detect micro-areal patterns.

In the following section we introduce our data and provide a definition of what we call 'non-canonical SAY'; in Section 3 we discuss the different forms and functions that we were able to identify in the languages we studied, while in Section 4 we attempt to discern patterns in the multitude of data points with the help of exploratory

[2] Both tokens in Eastern Ėven were obtained in one location in the east of the Sakha Republic from two speakers with close personal connections to Sakha speakers. Individual contact-induced use of these constructions is thus highly probable. The eastern Ėven dialect spoken on Kamchatka, out of range of possible Sakha contact influence, shows no occurrences of non-canonical SAY whatsoever.

[3] Non-canonical SAY in Siberia has mostly been neglected in the literature. SAY in Mongolic and Turkic has been mentioned in many publications, but the Tungusic languages are less well described: for North Tungusic, apart from short sketches of non-canonical SAY in Evenki (Brodskaja 1987, 1988: 48-9, 72-3), there is one mention of a special use of SAY in Ėven (Malchukov 2008: 326-7); for South Tungusic, non-canonical SAY is mentioned only in comprehensive grammars of Udihe (Nikolaeva & Tolskaya 2001: 461ff, 662ff.) and Nanai (Avrorin 1961: 275-6). Non-canonical SAY in small language families and isolates has up to now received practically no attention at all: we are only aware of brief mentions of subordinating SAY in Kolyma Yukaghir by Nikolaeva (2005: 312) and of the quotative morpheme *-vu-* (< SAY) in Nivkh in the standard grammar by Panfilov (1965: 122-123). No comparative work has yet been done.

Correspondence Analysis. Finally, Section 5 discusses the areal and historical factors that have shaped the distribution of non-canonical SAY in Siberia.

2. Preliminaries

2.1. Data

For the purposes of this paper, 'Siberia' is defined as the territory between the Ural Mountains to the west and the Pacific Ocean/Sea of Okhotsk to the east, and between the Arctic Sea in the north and roughly the border between Russia and China to the south (Figure 1). These boundaries are obviously fairly arbitrary: while there is a natural barrier to the diffusion of linguistic features to the north and east, the western boundary, even though represented by a mountain range, is definitely not impermeable, and this holds even less for the southern boundary, which is a political border that had little meaning in the past. It is thus possible that the phenomenon of non-canonical SAY in Siberia has historical connections with similar structures in East and Southeast Asia (e.g. Matisoff 1991, Bisang 1992: 49, Chappell 2008), the Himalayas and South Asia (e.g. Ebert 1986, 1991, Saxena 1988, 1995, Noonan 2006, Genetti 2011), and the Turkic and Turkic-influenced region spreading from Central Asia to the Balkans (e.g. Pokrovskaja 1978: 156ff, Johanson 2002: 137, Erdal 2004: 488ff, Khanina 2007, Straughn 2008).

Figure 1: Distribution of languages included in the sample

The primary data on which we based our study is natural discourse (mostly narratives), supplemented where possible by grammatical descriptions (see Figure 1 for the geographical location of the languages and Appendix 1 for details on the corpora). The sample consists of both unpublished corpora of interlinearised oral narratives from our and our colleagues' field data as well as published collections of texts, even though we are well aware that such texts may be heavily edited. We included the Ob-Ugric languages Khanty and Mansi, the Samoyedic languages Nganasan and Enets, the Mongolic languages North Mongolian and Buryat, the Turkic languages Tuvan, Shor, Sakha (Yakut), and Dolgan, the Tungusic languages Evenki (both western Evenki and eastern Evenki dialects), Ėven (both western Ėven and eastern Ėven dialects), Negidal, Nanai, and Udihe, the Chukotko-Kamchatkan languages Koryak and Alutor, as well as Ket, Kolyma Yukaghir, and the isolate Nivkh. We scrutinized the texts for forms of the generic verb of speech, which we had previously identified from dictionaries and grammatical descriptions, and categorized the different tokens of SAY as finite verbs, canonical use of non-finite forms, or non-canonical SAY. In the latter case, they were further classified according to their function. Although most of the texts we used are folklore, in the oral corpora of Sakha and Dolgan other types of narrative discourse prevail. In order to control for the effect of genre on our results, we complemented these with published folklore texts comparable to those used for the Mongolic languages, Tuvan, and Evenki. Our Western Ėven field data contained both non-folklore discourse and a sufficiently large amount of folklore texts for us to be able to investigate the occurrence of non-canonical SAY separately in both genres (see discussion in Section 5.4). The size of the corpora ranges from ca. 3,500 words (Ket) to ca. 48,300 words (Western Ėven life histories), depending on the data available and on the presence of non-canonical SAY in the language.

2.2. The matter of investigation

2.2.1. Lexical properties of SAY in Siberia

It has been argued that the diachronic developments which result in non-canonical uses of SAY do not necessarily have their source in generic speech verbs, but may originate in various types of pro-verbs, similative expressions, deictic pronouns, etc.

(Güldemann 2008: 264ff). In order to ensure that we are basing our study on truly comparable structures, we first need to show that in all the languages in our sample non-canonical SAY stems from a similar source, namely a generic speech verb. A verb can be considered a fully-fledged generic verb of speech if it appears in more syntactic contexts than merely the introduction of quotes, if it can host the full range of inflectional affixes and serve as a basis for productive derivation, and if it keeps its utterance meaning across syntactic contexts (Güldemann 2008: 271).

The languages of Siberia fulfil these criteria. The use of the relevant verb with speech-content complements other than direct quote is universally attested in our corpora and/or in dictionaries, as in (2). Furthermore, most of the SAY verbs in our corpora can be used without a speech-content complement in order to denote a general involvement in a speech act, as in (3).

(2) Western Ėven (DM2007 – Raven_ZKM2_021)
Eliwum *goːn-če!*
truth.ACC say-PF.PTC
'He was telling the truth!'

(3) Kolyma Yukaghir (Nikolaeva 2004: 34.6)
... *tabaːq* *oːžə-t* *lʼə-t* *met-kələ* *moŋ-ŋi-tə-j*
tobacco drink-IPF.CVB do-IPF.CVB I-ACC say-PL-FUT-3
'... they were smoking etc., and they would talk to me.'

There is no evidence that any of the SAY verbs in our sample is morphologically defective. We take this as an indication that we are not dealing with formulaic quote markers, possibly derived from non-SAY sources, but with regular speech act verbs. Also, lexemes derived from these verbs are invariably related to speech, e.g. Sakha *dieččі* [say.NLZR] 'speaker', Evenki *gu-ge* [say-NLZR] 'talkative', Ėven *goːr-mčin* [say-NLZR] 'idiom', etc. Finally, in all the languages in our sample, SAY is clearly distinct from other potential sources of non-canonical structures, such as general activity verbs (DO), as can be seen from the comparison of *xi-/xe-* (DO) vs. *ge-* (SAY) in North Mongolian and Buryat; *kylyr-/qyl-* (DO) vs. *de-/te-* (SAY) in Tuvan and Shor; *gïn-* (DO) vs. *die-* (SAY) in Sakha and Dolgan; *nek-/ńeke-* (DO) vs. *goːn/gun-* (SAY) in Ėven and Evenki (with cognate forms in Nanai and Udihe); *lʼə-* (DO) vs. *mon-* (SAY)

in Kolyma Yukaghir; *nə-* (DO) vs. *fur-* & *it-* (SAY) in Nivkh; and *ʃeda-* (DO) vs. *man-* SAY in Enets. The source of the structures we label non-canonical (SAY) is thus clearly a generic verb of speech in all the languages of Siberia, and our diachronic endeavour based on comparative evidence is justified.

2.2.2. Non-canonical use vs. grammaticalisation

The Yukaghir same-subject converb *monut* 'saying' in example (1) clearly has a function quite distant from the primary meaning of the speech verb: rather than denoting a speech act, it is used to mark a purpose clause. This extended use would commonly be labelled *grammaticalised SAY*. However, although many of the extended uses we find amongst the languages of Siberia might be defined as grammaticalised, a narrow focus on fully grammaticalised uses would exclude a large variety of conventionalised, but still not fully grammaticalised functions, as will be discussed in this section. Consider first examples (4) and (5), both from W.Even:

(4) Western Even (DM2007 - Creation_Animals.S.Golikova.4.030)
«*E-dʒi merget-te* (...) », ***go:niken*** *o:ŋat-u bel-e-hn-i-n.*
NEG-IMP.2SG think-NEG.CVB say.SIM.CVB nose-ACC touch-0-LIM-PST-3SG
'«Don't you worry (...) », he said and touched my nose.'

(5) Western Even (DM2009 - Reindeer_masti_GNM_4_1.263)
Buju-hel (...) *ere ma:-dʒi-ndi dʒeb-de-ji* ***go:niken.***
wild.reindeer-PL only kill-FUT-2SG eat-PURP.CVB-REFL.SG say.SIM.CVB
'Wild reindeer (…) you'd kill them only in order to eat them.'

Sentence (4) contains an instance of what we call **canonical SAY**. Canonical SAY is every form of SAY which is semantically and syntactically well-formed on a purely compositional reading. Thus, the same subject converb *go:niken* in (4) is canonical because the sentence is well-formed (a) when *go:niken* is interpreted as denoting a voluntary production of meaningful sounds, and (b) when *go:niken* is interpreted as syntactically dependent on the matrix predicate, with which it shares the subject. In other words, the term *canonical* here refers to the possibility to derive the meaning of

a sentence resorting only to the literal meaning of SAY and to the general syntactic rules for the construction in which it is used.

The same converbal form in (5) is an instance of **grammaticalised SAY**, i.e. a form of SAY in which the word class of SAY has been changed. There are various symptoms of such a change, not all of which have to co-occur: (a) syntactic and semantic ill-formedness on a purely compositional reading, (b) changes in inflection, and (c) phonological erosion. While the form *go:niken* in (5) does not display any changes in inflection or phonological erosion, it displays both types of ill-formedness. On a compositional reading, it would be syntactically ill-formed, since the verb SAY does not take converbs (here: purposive converb *ʤebdeji*) as its complements, and semantically ill-formed, since the situation described does not include the action of producing meaningful sounds. This implies that *go:niken* in (5) must be treated as a lexeme which is distinct from canonical SAY, as a complementiser introducing purpose clauses, even though it has not undergone inflectional or phonological changes. In other words, not all the symptoms have to be present in order to diagnose a full change in word class.

The phenomena of (b) inflectional change and (c) phonological erosion are also represented in some languages of Siberia. In Sakha, the perfective converb of SAY, *dien*, can receive nominal case suffixes, a morphological property that can be explained only by the assumption that in this construction it has lost its verbal character and grammaticalised into another word class (6). Phonological erosion is found in Nanai: the erstwhile same-subject converb of SAY *umi* (cf. Avrorin 1961: 276 for discussion) has turned into a clitic quotative particle *=Em* (7).

(6) Sakha (BP2002 – Efmy_392)
Onton Valja-lar Valerka-lar Lenskaj_Ostuolbi ***die-ŋ-ŋe***
then V.-PL V.-PL Lena_Pillars say-PF.CVB-DAT
bar-bit-tara.
go-PST.PTC-3PL
'Then Valja and Valerka went to the so-called Lena Pillars.'

(7) Nanai (Avrorin 1961: 275; our glosses)
*Ńoani min-či daŋsa-wa bu-ru=**əm** un-ki-ni.*
3SG 1SG-ALL book-ACC give-IMP.2SG=SAY.BND say-PST-3SG

‘He said: «Give me the book».’

The uses of SAY illustrated in (5)-(7) are easy to classify as grammaticalised and distinct from canonical SAY. However, our corpora also contain instances of SAY which, on the one hand, do not seem to be canonical in the narrow sense of the word, but do not necessarily imply a full word class change on the other. Sentence (8) is a good example of the ambiguous status of these kinds of examples:

(8) Sakha (BP2002 – MalA_98)

Elbeχ *oγo-nu* *kïaj-an* *kör-üö* *huoχ-pun* ***die-n***
many child-ACC be.able-PF.CVB see-FUT.3SG NEG-PRED.1SG say-PF.CVB
akka:s-ta-n-an *olor-obun.*
refusal-VR-REFL-PF.CVB sit-PRS.1SG
‘Saying that I will not be able to look after many children, I refuse.’
‘Since I will not be able to look after many children, I refuse.’

In (8), the converb *dien* displays the syntactic properties of a converb of SAY, but its semantic contribution is different from the canonical use of SAY in many languages, as no action of producing meaningful sounds need be involved. It can still be interpreted as canonical, since the lexical content of SAY in Sakha also covers internal monologue (i.e. thoughts, cf. Slepcov 1972: 114; see also Section 3.3.2).[4] However, such internal monologue clauses are one of the major conventionalised ways of expressing subjective causal relationships between states of affairs in Sakha. Thus, (8) is essentially ambiguous between a canonical interpretation of a speech act and one of *dien* functioning as subordinator for reason clauses (see especially Chisarik & van der Wurff 2003 on ‘inner speech’ and reason clauses). Thus, (8) cannot be simply relegated to the category of canonical SAY, even though the use of SAY in such constructions is not fully grammaticalised.

What (8) illustrates is that there is a class of uses of SAY which is conventionalised without being grammaticalised. This type of less than fully compositional structure, labelled *use pattern* by Heine & Kuteva (2005: 44), often represents an early stage of grammaticalisation and can be copied from one language to another, as any

[4] This broad use of SAY is widely attested in the world’s languages, e.g. in Quechua (Adelaar 1990), Usan (Reesink 1993), Manambu (Aikhenvald 2008), Bengali (Chisarik & van der Wurff 2003), etc.

conventionalised construction can. In order to capture both of these types of uses of SAY, we here define a category of **non-canonical SAY**, intended to comprise all instances of SAY which deviate from the fully transparent construal, regardless of the degree and type of deviation. In this way, the category of non-canonical SAY investigated here includes both fully grammaticalised and merely conventionalised uses of SAY and provides for a much richer basis of comparison within and across languages.

2.3. Non-canonical SAY in Turkic and Mongolic

In contrast to the bulk of the languages included in our sample, for which nothing or not much is known about non-canonical SAY, the Turkic and Mongolic languages are well known to make heavy use of this feature. It is attested in a wide array of functions in the earliest runiform records of Turkic from the 7-10th century as well as of Old Uyghur from the 9-13th century, and these uses of SAY have persisted in the family until today (cf. Table 2). Interestingly, even though the basic constructions are identical across Turkic, there has always been some variation in the forms used: Erdal (2004: 504) notes that non-canonical SAY in Old Uyghur appears as the anterior converb *tep*, whereas runiform inscriptions use a different converbal form, *teyin*. Most modern Turkic languages use a converb with anteriority meaning (cf. Tuvan non-canonical *dep*, Shor *tep*, Tatar *dip*, Balkar *dep,* Uzbek *deb*, Sakha and Dolgan *dien*), but Turkish uses the simultaneous converb *diye* (Johanson 1998a: 48, 1998b: 117).

Table 2: Overview of non-canonical SAY in Turkic languages

Language	Reference	Functions
Old Turkic	Nedeljaev et al. 1969: 545 Erdal 2004: 488ff.	quote marker complementiser purpose and reason adjunct name with verb of calling
Chuvash	Khanina MS	quote marker complement of cognition/emotion/ perception verbs purpose and reason adjunct name with verb of calling
Gagauz	Pokrovskaja 1978: 156ff.	
Turkish	Kornfilt 1997: 2ff, Göksel & Kerslake 2005: 175	
Uzbek	Straughn 2008	

Tatar	Khanina 2007	name in apposition
Turkmen	Clark 1998: 456	

The situation in the Mongolic languages seems to be similar. The use of SAY as a quote marker with non-generic verbs of speech is attested in the earliest extant records of Middle Mongolian (13th century, Slater 2003: 310); further extended uses of SAY are attested in Classical Mongolian and all extant Mongolic languages (cf. Table 3). Two converbal forms seem to dominate: the modal converb in *-n/-v* (*geen/kemen/gamav*) in Middle and Classical Mongolian, and different forms of the simultaneous converb in *-ž(V)* in the modern languages (Khalkha *gež,* Kalmyk *giž*, Dagur *gaji*, Mangghuer *geji*, Bao'an *=tɕə*, cf. Fried 2010: 294); furthermore, participial forms in non-canonical use are also widely attested (e.g. in Kalmyk – Muniev 1977: 141-5, in Mangghuer – Slater 2003: 303ff, etc.).

What these short summaries show is that the use of converbs of SAY as quotative and complement marker with at least cognition verbs, as well as to mark names in apposition and with a verb of calling, is probably an inherited feature in the Turkic and Mongolic languages.

Table 3: Overview of non-canonical SAY in Mongolic languages

Language	**Reference**	**Functions**
Middle Mongolian	Slater 2003: 310	quote marker
Classical Mongolian	Poppe 1937: 194ff, Lessing 1960: 451, Janhunen 2003: 54	quote marker complement cognition verbs name with verb of calling name in apposition auxiliary
Khalkha	Binnick 1979: 70, 101 Svantesson 2003: 173ff.	
Chakhar	Sechenbaater 2003: 152ff.	
Kalmyk	Muniev 1977: 141ff.	
Dagur	Martin 1961: 150 s.v. gaji	
Bao'an	Fried 2010: 294ff.	
Mangghuer	Slater 2003: 298ff.	
Mongghul	Sanžeev 1964: 134	name in apposition

3. Non-canonical SAY in Siberia: Frequency, forms and functions

3.1. Frequency

With the exception of the Yeniseic, Ob-Ugric, and the Chukotko-Kamchatkan languages, non-canonical SAY is present in all the language families of Siberia, albeit with varying frequency, as summarized in Table 4.

Table 4: Frequency of non-canonical SAY in the languages of Siberia, ordered by rank

Language family	Language/dialect	# tokens SAY	non-canonical SAY	
			%	# tokens
Yeniseic	Ket	49	0	0
Samoyedic	Nganasan	73	0	0
Ob-Ugric	Khanty	243	0	0
	Mansi	106	0	0
Chukotko-Kamchatkan	Koryak	201	0	0
	Alutor	46	0	0
North Tungusic	Negidal	165	0	0
	Western Evenki	476	0	0
	Eastern Ėven	608	0.3[5]	2
Isolate	Nivkh	74	8.1	6
North Tungusic	Eastern Evenki	529	12.5	66
Yukaghir	Kolyma Yukaghir	817	13.2	108
South Tungusic	Nanai	249	13.7	34
North Siberian Turkic	Dolgan (life histories)	372	14.0	52
	Dolgan (folklore)	201	15.9	32
South Tungusic	Udihe	278	16.9[6]	47

[5] As mentioned in footnote 1, these two tokens of non-canonical SAY are probably the result of individual contact with Sakha and not characteristic of the dialect as a whole.

[6] Nivkh and Udihe have two instead of one basic verb of saying (*it-* and *fur*, and *gun-* and *dian-*, respectively). In order to provide for comparability with other languages, we have included tokens of both verbs into the final number of tokens of SAY (Nivkh: 14 *fur-* + 60 *it-*, Udihe: 146 *gun-* + 132 *dian-*), even though only *fur-* and *gun-* are used non-canonically.

Samoyedic	Enets	122	20.5	25
North Tungusic	W.Ėven (folklore)	263	21.3	56
North Siberian Turkic	Sakha (folklore)	272	27.6	75
North Tungusic	W.Ėven (life histories)	933	34.4	321
South Siberian Turkic	Shor	180	40.4	71
South Siberian Turkic	Tuvan	389	48.1	187
Mongolic	Buryat	599	50.8	304
Mongolic	North Mongolian	362	56.6	205
North Siberian Turkic	Sakha (life histories)	659	66.4	438

In terms of frequency, three groups of languages in Siberia are discernible. The first group comprises languages that completely lack any non-canonical use of SAY: the Yeniseic, Chukotko-Kamchatkan and Uralic languages with the exception of Enets, as well as some North Tungusic lects. The lack of non-canonical SAY in the Yeniseic, Uralic, and Chukotko-Kamchatkan families as a whole is confirmed by its absence in languages related to those investigated here: there is no mention in descriptive studies and no textual evidence known to us for the presence of non-canonical SAY in any of the Yeniseic languages (cf. especially Werner 1997: 368ff), nor is this feature known in the Samoyedic languages Kamass (G. Klumpp, p.c.) or Nenets (I. Nikolaeva, p.c.), or the Chukotko-Kamchatkan language Chukchi (M. Dunn, p.c.). Apart from Enets, the only Uralic languages that appear to have non-canonical SAY are those that have been in close contact with Turkic languages: the Volgaic language Mari and the Permic language Udmurt (G. Klumpp, p.c; cf. Serdobol'skaja & Toldova 2006).

The second group is comprised of languages in which between 8 and 34% of all tokens of SAY are used non-canonically, while the languages of the third group make heavy use of non-canonical SAY, which comprises more than 40% of all the tokens of SAY. Interestingly, there are fairly large differences in frequency of non-canonical SAY between the life history and folklore corpora of Sakha and Western Ėven (though not of Dolgan), with the life history data making far heavier use of this feature than

what is found in folklore data. We discuss these differences between the genres in Section 5.4 below.

The frequency pattern of non-canonical SAY corresponds roughly to the geographical distribution of the languages/dialects (cf. Figure 2): the strongest concentration of non-canonical SAY is observed in the Mongolic and Turkic languages of the south (Mongolic and South Turkic) and northeast (Sakha life histories). Medium frequencies are found in the southeast (Nivkh, South Tungusic), northeast (Sakha folklore, W.Ėven, E.Evenki, and K.Yukaghir), and in the northwest, on the Taimyr peninsula (Dolgan, Enets). However, the presence of SAY in the northwest is explainable at least in part by an accident of history, since Dolgan is a very recent immigrant to the area. Non-canonical SAY is absent in the languages spoken at the periphery of the area: Ob-Ugric, Ket, Nganasan, and W.Evenki in the west/northwest, and Chukotko-Kamchatkan and E.Ėven in the far northeast. Enets (spoken on the Taimyr in the vicinity of Nenets, Nganasan, and Dolgan) and Negidal (spoken in the Far East in the vicinity of Nivkh, Udihe, Nanai, and Eastern Evenki) are unexpected outliers within their language families: Enets has a medium frequency of non-canonical SAY, which is completely absent in other Samoyedic languages, while Negidal lacks non-canonical SAY entirely, even though it is found at varying frequencies in other Tungusic languages. In what follows, we concentrate only on those languages in which non-canonical SAY is attested.

Figure 2: Frequency distribution of non-canonical SAY in Siberian languages, created with the WALS Interactive Reference Tool (Bibiko 2005)

3.2. Forms

As can be seen in Table 5, most languages of the sample that have non-canonical SAY use at least two different forms, with only Dolgan and Nanai employing a single form. On the other end of the scale, in Buryat there are ten forms that occur in non-canonical functions of SAY (cf. Skribnik 1987: 33). In the table, the predominantly used forms are marked in bold; this clearly shows that, when several forms are used for non-canonical SAY, one of them is usually the dominant one, occurring with high frequency and in a wide range of functions, while the other forms might be restricted to individual uses.

Table 5: Forms of SAY used with non-canonical functions in languages of Siberia

	Converbs	Participles	Particles	Bound morphs	Finite
Enets	SS.COND *mab*		*man'*		
North Mongolian	**SS.IPF *gedž***	HAB *gedeg*			various forms
	SS.PF *geet*	PF *ges(e)ŋ*			PRS.EMP *genee*
	SS.TERM *getel(e)*	PRS.ACT *gektši*			
Buryat	**SS.IPF *geže***	HAB *gedeg*			various forms
	SS.PF *geed*	FUT *gehe*			
	SS.COND *gebel*	FUT-DAT *gehede*			
		PRS *geese*			
		PST *gehen*			
		REC.PST *gee*			
Tuvan	**SS.CNJ *dep***	PRS *deer*			various forms
	SS.SEQ *deeš*	PRS-DAT *deerge*			
		PST *deen*			
Shor	**SS.IPF *tep***	PST *teen*			
	SS.COND *teze*				
Sakha	**SS.PF *dien***	MDL-DAT *dieteχχe*			
	SS.IPF *di:*	NEG.PRS-DAT *diebekke*			
Dolgan	SS.PF *dien*				
Kolyma Yukaghir	**SS.IPF *monut***				
	SS.PF *mo(nde)lle*				
	SS.ITER *monde*				
Western Even	**SS.SIM *go:niken***				
	SS.COND *go:mi*				
	SS.MUL *go:nteken*				

E. Evenki	SS.SIM ***gunne***	PRS *guneri*	*gun*		
Udihe		HAB *gune(i)*	*gum(u)*		
Nanai				clitic *=(E)m*	
Nivkh			*furu*	suffix *–vu–*	

Overall, the most frequent dominant forms of non-canonical SAY are same-subject converbs (cf. Figure 3), as already illustrated in examples (1), (5), (6), and (8). In the Mongolic and North Tungusic languages as well as Kolyma Yukaghir the predominantly used converbs are of the simultaneous type (also called imperfective), i.e. they are interpreted as broadly contemporaneous with the time of the action expressed by the main predication. Various iterative converbs and the conditional converb in Ėven also fall under this general label of simultaneity. Converbs with an anteriority interpretation (conjoining, perfective, sequential, etc.) are the dominant form used in the Turkic languages as well as representing a minor subtype of non-canonical SAY in the Mongolic languages and Yukaghir. Converbs are not used at all only in the southeast, in Nanai, Udihe, and Nivkh.

Participles with a non-canonical function are used by Mongolic (9) and some Turkic languages, as well as by two Tungusic languages (Eastern Evenki and Udihe).

(9) North Mongolian (NMV 1974: p.14: 28.44; our glosses)
Buudää Šarw ***ge-dek*** *xün-iig tani-xa xüŋ bae-n-uu?*
B. Sh. say-HAB.PTC person-ACC know-FUT.PTC person COP-PRS-Q
'Is there (any) person who knows a man called Budä Sharw?'

Particles derived from SAY are found in the southeast, in Nivkh (*furu*, from *fur-* 'say'; Panfilov 1965: 123), and Udihe (*gum(u)*, from a passive form of *gun-* 'say'; Nikolaeva & Tolskaya 2001: 461), further north in E.Evenki (*gun*, from *gun-* 'say'), and, on the northwestern margin of the area, in Enets (*man^{j}* from *man-* 'say'). Fully grammaticalised bound morphemes are attested only in the southeast, in Nivkh (suffix *-vu-*, probably from *fur-* 'say', Mattissen 2008: 113, cf. ex. (10) below) and in Nanai (Section 2.2.2, ex. (7)).

(10) Nivkh (Mattissen 2008: 112, ex. 85)

If *přə-r* *p'-ŋafq-ax* *osqa-vil-**vu**-r*

3SG come-CVB.3SG REFL-friend-CSE cowardly-big-SAY.BND-CVB.3SG

it-t'

say-IND

'He came and said that his comrade is a coward.'

Finite forms of SAY are usually not used non-canonically, except in their capacity as auxiliaries in Mongolic and Tuvan in the south (Section 3.3.9).

Figure 3: Distribution of major forms of non-canonical SAY in Siberia, created with the WALS Interactive Reference Tool (Bibiko 2005)

black symbols: use of converbs, white symbols: converbs not used

circles: only converbs; squares: + participles; diamonds: + participles + particles; triangles: bound forms; large circle marks languages that use finite forms

3.3. Functions

In this section, we discuss the functions of non-canonical SAY, most of which have been identified in other languages and areas in previous work (cf. Güldemann 2008 and Chappel 2008: 49). The sections contain only short descriptions with illustrative examples and overviews over the areal and genealogical distribution of certain functions; detailed information on the frequency with which each function is attested in the languages of the sample is included in Appendix 2.

Practically all functions of non-canonical SAY in our sample can be reduced to one common semantic/pragmatic feature, which, following Güldemann (2008: 6), we label

dissociation and define as the "representation of a spoken or mental text from which the speaker distances him/herself by indicating that it is produced by a source of consciousness in a pragmatic and deictic setting different from that of the immediate discourse". Since the functions of non-canonical SAY in the languages of Siberia are derived from the same source, they represent a continuum with blurred demarcation lines rather than a set of discrete entities. In spite of this, we treat them as separate independent units. The main reason for this is that clearly defined data points, i.e. discrete functions, facilitate cross-linguistic comparison; in addition, any scientific approach to continua requires the individual points of the continuum to be identifiable, and any such identification requires a certain amount of discreteness in the method. The assignment of individual items to separate functions on the continuum was carried out using unified criteria for all languages (described below), to ensure comparability. The functions described in the following subsections are grouped on the basis of syntactic, semantic, and pragmatic similarity for ease of presentation, but not all the functions grouped in this way necessarily share a common diachronic source. In the statistical analysis discussed in Section 4, each function (labelled with an abbreviation in the text) was treated as a separate data point.

Before proceeding to the description of individual functions of non-canonical SAY, one further clarification is in order. Attempts have often been made to construe a universal grammaticalisation path that SAY has to follow in its development from a generic verb of speech to quotative marker, complementiser, comparative marker, etc. (e.g. Ebert 1991, Saxena 1995, Heine & Kuteva 2002: 261ff). We shall largely ignore attempts of this kind in this study, mainly because the exact shape of the universal grammaticalisation path of SAY is far from being a settled issue yet, and it is questionable whether such a path exists at all (see Güldemann 2008: 267ff. for a detailed critique). We shall, however, comment on the relationships between different functions (see especially Section 3.3.11), mainly in order to shed light on the larger groupings to which they belong and to highlight the interconnectedness of the whole non-canonical-SAY complex.

3.3.1. Marker of direct speech

The most frequent use of non-canonical SAY in our data is as a marker of direct speech. The general form of the construction includes non-canonical SAY attached to

a quote in order to signal its dissociated status (see above). The main criteria for classifying an instance of non-canonical SAY as a quotative marker are: (a) the clause that SAY is adjoined to refers not to the extralinguistic world, but to verbal entities, and (b) there is no deictic shift within the quote. Tokens displaying incomplete or complete shift in deixis are classified as indirect speech (Section 3.3.2; cf. Güldemann 2008: 8ff. on the gradual nature of deictic shift in reported speech).

The structure of the quotative construction in Siberia is quite uniform: the finite clause encoding a quote is followed by non-canonical SAY. Depending on the syntactic embedding of this structure and the way the quote is encoded, a number of subtypes can be differentiated. A quote can be introduced by a different verb of speech (**Qu_DiffV**); in this case, the matrix verb is as a rule more specific than the generic verb of speech, denoting the manner of saying or the type of speech act – question, answer, exclamation, etc. – produced by saying (11). In some languages of Siberia, quotes can be introduced by the same generic verb of speech from which non-canonical SAY is derived (**Qu_SameV**, (12)), and, in similar constructions, by nouns denoting a speech event (**Qu_Noun**; (13)). Furthermore, non-canonical SAY can function as the sole marker of direct speech, with no other finite matrix verb introducing the quote (**Qu_NoV**; (14)). Finally, in a few languages a quote can be referred to with an anaphoric pronominal adverb functioning as a modifier of non-canonical SAY (15). This structure can be embedded in any of the above-mentioned syntactic contexts; given the small number of instances, we subsume them all under one category (**Pron_Qu**).

(11) Buryat (BNS 2000: p.168, 36.149; our glosses)
«Jexel gojo baj-na daa!» ***ge-že*** *lama-nar xarjuusa-ba*
very beautiful be-PRS PTL say-IPF.CVB lama-PL answer-PST(3)
'«It was very beautiful (there)!» answered the lamas.'

(12) Tuvan (TNS 1994: p.306, 9.6; our glosses)
«Kandyg janzylyg süt-te-p iž-er sen, ačaj?»
how milk-VR-CNJ.CVB drink-NPST 2SG father
de-p, *uruu ynča de-en.*
say-CNJ.CVB daughter.3SG like.this say-PST(3SG)
'«Father, how will you get milk?», his daughter said.'

(13) Buryat (BNS 2000: p.160, 36.4; our glosses)

«Tede-ner (...) *iime šadal-taj sese merge=šüü=l»*

that-PL like.that strength-PROP wise clever=PTL=PTL

ge-že *olon zon-oj dunda tunxag tara-ba*

say-IPF.CVB many people-GEN middle rumour spread-PST(3)

'«They (...) are so very clever», such a rumour went amongst the people.'

(14) Nanai (Sofia Oskolskaja field data 2009)

Ča-du dami-su ***bi=əm*** *gə*

that-DAT grandfather-2PL be=SAY.BND PTL

'«There is your grandfather».'

(15) Western Ėven (DM2007 - S.Golikova.1.079)

Tačịn go:niken *ukče:n-di-n bi: o:hirep abaga-w.*

thus say.SIM.CVB tell-PST-3SG 1SG ancient grandfather-1SG

'Saying like that my old grandfather used to tell.'

Enets in the northwest and Nivkh in the southeast strongly diverge from the rest of the SAY area with respect to non-canonical SAY as a quote marker. Enets is the only language in the sample making use of non-canonical SAY where the quotative construction is entirely lacking, and Nivkh is conspicuous for its limited use: the morpheme *-vu-* is restricted to marking direct speech with the verb *it-* 'say' (cf. Panfilov 1965, Krejnovič 1979, Mattissen 2008; ex. (10) above).

In the Turkic and Mongolic languages, Qu_DiffV is obligatory: this usage was practically exceptionless in Old Turkic (Erdal 2004: 506), and is still obligatory in such geographically non-contiguous languages as Turkish (Kornfilt 1997: 2, Göksel & Kerslake 2005: 352), Altai (Čeremisina 1987: 7), and, in our sample, in Sakha. Similar to Turkic, the obligatoriness of Qu_DiffV seems to be an ancient feature in the Central Mongolic branch (Buryat – Skribnik 1987: 34, Khalkha – Binnick 1979: 100).

This is reflected in the extremely high frequency with which this function occurs in the languages of South Siberia investigated here (N.Mongolian, Buryat, and Tuvan).[7]

The syntactic status of quotes and their matrix clauses has been controversially debated. The standard view, according to which the quote is a complement of the verb of saying, has been challenged recently (see e.g. McGregor 1994, Collins & Branigan 1997, Suñer 2000, Güldemann 2008: 224ff), and analyses in terms of separate clauses and adjunction have been proposed. The available syntactic evidence from Siberia is not unequivocal. In some languages, quotes do not seem to function as complements. The case in point is Kolyma Yukaghir, in which the SAY verb *mon-* always carries the intransitive set of agreement suffixes when introducing quotes, indicating that the latter cannot be considered its direct objects. On the other hand, in most Turkic and Mongolic languages, quotes are subject to rigid selectional restrictions, such that only the general verb of utterance (SAY, i.e. *ge-* in Mongolic and *de-/te-* in Turkic) can combine with them them directly; with all other verbs of saying, an additional element (usually non-canonical SAY) is needed. Since adjuncts and coordinate clauses are adjoined at the clause level, not at the predicate level, if quotes were adjuncts/independent clauses, they should not be sensitive to the matrix verb. As they obviously are sensitive to the matrix verb, quotes in Turkic and Mongolic can be considered properly subcategorised objects of the verbs of saying, i.e. complements. The languages of Siberia thus seem to differ with respect to the syntactic status of the quote – it can display properties of a proper complement, as in Turkic and Mongolic, or it can be a separate clause, independent of the SAY clause, as in Yukaghir. One could argue that in the former group of languages, non-canonical SAY has a double function as a marker of dissociated speech and complementiser, whereas in the second group it just marks dissociated speech.[8]

[7] That genre also plays a role in the frequency distribution of this function is demonstrated by the separate Sakha and Dolgan corpora, where Qu_DiffV is more frequent in the folklore texts than in the life histories (cf. Section 5.4 for an explanation).

[8] For ease of reference, we will keep on using the term 'matrix verb' and 'matrix clause' for both types of languages.

3.3.2. Complementiser

The dominant means of complementation in the languages of northern Eurasia is nominalisation (cf. e.g. Kornfilt 2001, Anderson 2006); the major competing strategy are finite clauses introduced with non-canonical SAY, attested with verbs of saying, cognition, emotion and perception. As mentioned in Section 2.2.2, in many languages of Siberia the denotation of generic SAY covers not only voluntary production of sounds, but also internal monologues, which enables them to be broadly employed in the expression of internal awareness (Güldemann 2008: 422ff). This lexical property is responsible for the use of non-canonical SAY not only in subordinate structures with a verb of saying as the matrix verb, but also in those that include verbs of cognition, emotion, perception, etc, as shown in example (16).

(16) Western Ėven (BP2008, Stepanova_ZA_parents_06)

Ọ:n bi: ụnta-wa-n haŋan-ʤị-m ***go:niken*** *merget-ti-w.*

how 1SG fur.boot-ACC-3SG sew-FUT-1SG say.SIM.CVB think-PST-1SG

'I thought about how I will sew fur boots.'

'I thought, saying «How will I sew fur boots».'

According to our criteria of non-canonicity, both possible readings of this sentence, the internal dialogue and the complementiser reading, are non-canonical and relevant for the present study. However, these sentences pose two kinds of problems, the practical problem of whether to classify a token of non-canonical SAY as a direct speech marker or a complementiser, and the theoretical problem of the syntactic status of the embedded clauses. The former issue was handled pragmatically by adhering to the following principles: if the internal-dialogue reading is impossible for semantic or pragmatic reasons, non-canonical SAY is classified as complementiser; if both readings are possible, we followed the judgments of our informants and their preferred translational equivalents, where possible.

The theoretical difficulty is similar to the one addressed in relation to quotes: is the clause followed by non-canonical SAY indeed subordinated to the matrix clause? This issue has been topical since Hooper & Thompson (1973), where it is argued that there are two types of clauses with complements, the canonical type (complementiser clause dependent on matrix clause), and the one with a reversed dependency relationship, where the 'matrix clause' functions as a parenthetical. Recent studies have shown that

that this distinction is reflected in the syntax and prosody of complex sentences (Boye & Harder 2007, Dehé & Wichmann 2010). This is of relevance for our purposes insofar as the status of the complement clause can reveal whether non-canonical SAY functions as a complementiser proper or merely as a marker of dissociation, similar to the situation with direct quotes. The criteria adduced for quotes are not applicable here: non-canonical SAY is not obligatory with verbs of cognition, emotion, etc., in any of the languages in our sample, and the formal markers of transitivity are not evenly distributed across different predicates. Nevertheless, there are indications that non-canonical SAY has at least some features of a complementiser. First, as illustrated in (17), its contribution to the sentence meaning goes beyond the pure marking of dissociation:

(17) Nanai (Avrorin 1961: 275; our glosses)
*Mergen tuj egdʒi dʒo-sal em boa-du bi:=**em***
hero thus many house-PL one place-DAT be.PTC.NPST=SAY.BND
xali=da ečie merčieni.
ever=EMP NEG think.PTC
'The hero never thought that there are so many houses in one place.'

The clause modified with non-canonical SAY is not dissociated, since there is no external consciousness that has produced its content (a proposition that a person does not know cannot be in their consciousness).[9] Thus, even though its origin is doubtless in dissociation, non-canonical SAY has come to play a syntactic rather than a semantic role in cases like (17). Furthermore, word order restrictions show that the matrix clause is not a parenthetical clause à la Hooper & Thompson (1973): there are no instances of the matrix clause inserted into the complement clause in our corpora, which should be possible if the former were parentheticals. This admittedly incomplete evidence shows that non-canonical SAY in the context of verbs of cognition and related predicates does not merely have a dissociative function and that the relationship between the complement clause and the main clause is not that of parataxis. Both of these features render the analysis of non-canonical SAY in the examples given in this

[9] The type of clause illustrated by (17), with a negated cognition, emotion or perception verb, is attested in all languages with the complementiser function of non-canonical SAY except for Dolgan and Shor.

section as complementiser plausible; similar to non-canonical SAY with quotes, this means that it functions both as dissociation marker and as a grammatical word.

SAY-complement clauses are formally identical to direct speech clauses, which is readily explained by their common origin (see above): a finite clause is followed by non-canonical SAY, and this structure is subordinated to a complement-taking matrix verb. In the languages of Siberia, non-canonical SAY is attested with complements of speech verbs, i.e. in indirect speech clauses (**SPCH**; (18); see Section 3.3.1 for the criterion used to distinguish indirect from direct speech), cognition verbs like 'know', 'find out', 'expect', etc. (**COGN**; (16), (19)), verbs of emotion such as 'fear', 'be glad', etc. (**EMOT**; (20)), and verbs of perception (**PERC**; (21)).

(18) Western Ėven (BP2008, StepZA_svatovstvo_21)

Tarkandụ *amm-ụ* *atịka-ŋ-ga-n* *ọ:-ʤi-m*

at.right.moment father-1SG old.woman-ALN-DES-3SG become-FUT-1SG

go:niken *go:n-če*

say.SIM.CVB say-PF.PTC

'She said right away that she will be my father's wife.'[10]

(19) Nanai (Sofia Oskolskaja field data 2009)

Murči-p *əm-bə* *ǯuər-bə* *elam-ba* *waa-o-ri=**əm***

think-1PL one-OBL two-OBL three-OBL kill-IMPS-PTC.NPST=SAY.BND

'We think that it would be good to kill one, two, three of them.'

(20) Eastern Evenki (Romanova & Myreeva 1964: 72, 7.15; our glosses)

Ewenki:-l *hulari:-r* *wa:-ʤiŋa:-tən* ***gun-ne-l*** *ŋe:le-ŋki-tən.*

Evenk-PL red-PL kill-FUT-3PL say-IPF.CVB-PL fear-PST-3PL

'The Evenks were afraid that the reds would kill them.'

(21) Shor (Irina Nevskaja, field data – Qara Qan 1061-2)

Tört *tajγa* *qaγ-yž-ypča* ***tep*** *kör-ze,*

four mountain hold-REC-PST say-SS.IPF.CVB see-COND.CVB

[10] Note that the deictic shift is only partial – while *ammụ* 'my father' takes the perspective of the narrator, as expected in indirect speech, *ọ:ʤịm* 'I will become' takes the perspective of the character, as expected in direct speech.

tört *alyp* *qarbaš-čytqan* *čer* *pol-tur*.
four warrior fight-IPF.PTC place be-INDIR.PST
'If you see that four mountains gathered, it is four warriors that are fighting there.'

Non-canonical SAY is used as complementiser almost universally, with the exception of the two languages on the fringe of the SAY area, Enets and Nivkh, which do not use conventionalised SAY in complement clauses at all. The Turkic languages have inherited at least COGN, EMOT, and PERC, the Mongolic languages at least COGN (cf. Section 2.3). COGN is also attested in all Tungusic languages that have non-canonical SAY (as well as in Manchu, cf. Gorelova 2002: 274), and in both Kolyma and Tundra Yukaghir, and is thus present in all the languages that have developed the complementiser-marking function.

3.3.3. Marker of adjunct clauses

Similar to complements, adjunct clauses in the languages of Siberia are normally encoded with non-finite verb forms (converbs and participles); finite adjunct clauses are as a rule introduced with non-canonical SAY. As in the case of complement clauses, the source of this type of non-canonical SAY is the lexical structure of SAY, which allows it to refer to internal monologues and denote states of internal awareness. For this reason, here, too, we find examples that are ambiguous between the internal speech interpretation and the adjunct interpretation, as in (8) above. In classifying these kinds of instances, we used the same criteria as with the complements (Section 3.3.2). Similar to complement-introducing non-canonical SAY, the status of SAY in adjunct clauses seems to be that of a proper subordinator: it displays symptoms of full grammaticalisation (cf. (1) and (5)), its word order properties are those of clause-final subordinators, the matrix clause is never inserted into the adjunct clause, and, at least in the case of purpose clauses, specific tense-mood-person agreement patterns are often grammaticalised (see below). Even though less widespread than quotative and complementiser functions, this use is very common in Siberia. However, the languages of the southeast (Udihe, Nanai, and Nivkh) do not use non-canonical SAY to introduce adjunct clauses at all.

Embedded finite clauses introduced with non-canonical SAY often stand in the semantic relation of reason to the main clause (**REAS**, (22)). This seems to be the most frequent interpretation of adjunct SAY-clauses across Siberia (see also (8)). In K.Yukaghir, even non-finite forms (converbs) expressing reason can be pleonastically introduced with non-canonical SAY, which bears witness to the advanced conventionalisation of this structure (23).

(22) North Mongolian (NMV 1974: p. 35, 31.206; our glosses)
Xosuu ixe džargal naer bol-dž, nojon-to
X. big joy feast become-PST(3) master-PROP
bol-dž ***ge-dž.***
become-PST(3) say-IPF.CVB
'It was a big joyous feast of Xosuu, because they have got a master.'

(23) Kolyma Yukaghir (Nikolaeva 2004: 15.45)
Tabun-gələ pulun-de: ta:t ibil'e:-l'əl tude terikə-gələ
that-ACC old.man-DIM thus cry-EV (3SG) he.POSS old.woman-ACC
kude-de-gə ***mon-u-t.***
kill-TR-DS.CVB say-0-IPF.CVB
'The old man started crying because they had killed his wife.'

Furthermore, adjunct clauses with non-canonical SAY often express purpose (**PURP**), as in (1) and (5). Purpose adjunct clauses differ from other subordinate clauses in the form of the embedded clause, which must be headed by a predicate with some kind of non-realis reference, frequently a verb in the future/non-past tense when the subject of the main verb and of the embedded clause are coreferential, and an imperative for different-subject situations. This is the case in most Turkic languages (cf. Bergelson & Kibrik 1995: 400ff. for Tuvan, and Pakendorf 2013: 263, 276f. for Sakha) and in Kolyma Yukaghir. In Shor, two types of imperative (Imperative 1 and 3) are used, the former for same-subject situations, the latter in different-subject cases (I. Nevskaja, p.c.).

(24) Sakha (BP2002 – Efmy_720)

*Min pervaj maːj-dïː bar-**iam** dʒe bar-dï-m **die-n***

1SG first May-VR.IPF.CVB go-FUT.1SG PTL go-PST-1SG say-PF.CVB

bat-tïm Bataːj-ga bïlïrïːn.

go-PST.PTC.1SG Batagaj-DAT last.year

'I went to Batagay last year to celebrate the first of May'

W.Ėven and E.Evenki have a special purpose construction with SAY: a non-finite form, the purposive converb in *-de-*, is optionally augmented with non-canonical SAY (cf. ex. (5)). In W.Ėven this is the sole purposive construction using non-canonical SAY, while for E.Evenki subordinate predicates in the future indicative and imperative mood are also attested (Brodskaja 1987: 62-3; see also Pakendorf 2013: 271f).

In addition to the fairly common functions as marker of reason and purpose clauses, non-canonical SAY is also found to introduce concessive (CONC) and conditional (COND) clauses. However, these functions are quite restricted and make use of different forms of SAY: concessive adjunct clauses are found only in Sakha (25), where they are introduced with a participial form of SAY, while only Enets uses the conditional converb of SAY to introduce conditional clauses (26).

(25) Sakha (BP2003 - Chir_189)

*Aɣïhuon haːs-taːχ kihi **die-teχ-χe**, (...) ïall-a*

eighty spring-PROP person say-MDL-DAT be.sick-IPF.CVB

ilik-pin

not.yet-PRED.1SG

'Even though I am eighty, (...) I haven't fallen ill yet.'

(26) Enets (Andrej Shluinsky field data, DJO_018)

***Mab** mɛzi, pogutʃ bɔɔ*

say.SS.COND.CVB wind fishing.net.FREQ.CVB bad

'If there is wind, it's bad to go fishing.'

3.3.4. Evidential

The development of evidential markers from non-canonical SAY in quotative functions is well documented cross-linguistically (see Aikhenvald 2004: 132ff, 271ff. and the references therein). In Siberian languages, the evidential use of SAY is closely related to the function of introducing quotes without a matrix clause (Qu_NoV, cf. 3.3.1). The structures employed in the evidential and Qu_NoV functions are identical – independent finite clause followed by non-canonical SAY, no matrix verb. The main difference lies in the semantic structure: while Qu_NoV marks the proposition as direct speech, evidential marking indicates the nature of the evidence for the given utterance. As noted in the literature (Güldemann 2008: 407ff), it is often difficult to differentiate between Qu_NoV and evidential functions. In less clear cases, we applied the criterion of the identifiability of the speaker/thinker: if it is possible to determine who said/thought the proposition denoted by the clause, we classified it as Qu_NoV; if not, it was counted as evidential.

Non-canonical SAY can signal that the encoded state of affairs is a piece of information for which the speaker has only second-hand evidence (hearsay evidential, **Evid_HSY**). In languages in which the exact semantic effect of this construction is assessable, this type of evidential meaning combines Aikhenvald's (2004: 64) categories of Assumption and Hearsay proper (e.g. Krejnovič 1979: 316ff. on Nivkh and Nikolaeva & Tolskaya 2001: 461ff. on Udihe). The translation 'they say' in the examples is an attempt to capture this effect:

(27) Nivkh (Shiraishi & Lok 2008, V.1.21-22; our glosses)
Ńi imŋ ətik χaj-nəta, χa-t' ***furu***.
1SG 3PL aunt be-HORT be.so-IND SAY.PTL
'I am supposed to be their aunt, that's how it is, they say.'

(28) Enets (Andrej Shluinsky field data, DJO_083)
Ibl^{j}ɛjgu-ɔn mɛuru-da ***man^{j}***
small-PROL.SG get.dark-FUT(3SG) SAY.PTL
'It will get dark in a bit, they say.'

Another type, perceptive/mirative evidential (**Evid_Perc/Mir**), is found in Udihe and Nanai. It marks states of affairs for which the speaker has first-hand perceptual

evidence. Explicit marking of the perceptual evidence is not obligatory; instead, this structure is mostly employed when a mirative effect is intended, i.e. when the witnessed state of affairs is unexpected (this is a cross-linguistically rare pattern: Aikhenvald 2004: 207ff. mentions only indirect, non-witnessed evidentials that have mirative effects).

(29) Nanai (Avrorin 1986: p.24, 1; our glosses)
*Mamaːčãː mapaːčaːn-či ičə-dʒiː-ni – mapaːčã=**m** bi-əsi!*
old.woman old.man-ALL see-PTC.NPST-3SG old.man=SAY.BND be-HAB
'The old woman looks at the old man – it is the old man!'

(30) Udihe (Nikolaeva et al. 2003, p.36, 5.79)
*Utadu emu(ge) meŋde waja-masi-n(i) **gune**, mamasa-ni tene*
then cradle with swim-DIV-3SG say.HAB.PTC wife-3SG CTR
*bie **(g)une***
be.PRS.HAB say.HAB.PTC
'Here was his wife swimming with the cradle (i.e. he saw her swimming).'

Two areal phenomena are conspicuous (see Appendix 2). First, in the extreme southeast and northwest of the SAY area, Nivkh and Enets stand out due to the high frequency of Evid_HSY, in both cases expressed with morphologically non-transparent particles. Second, Udihe and Nanai are the only languages with Evid_Perc/Mir, which is thus an obvious South Tungusic innovation.

3.3.5. Non-canonical SAY as a marker of metalinguistic use

In a number of Siberian languages, non-canonical SAY can mark elements in its scope as being metalinguistically used. It is employed to "signal a shift in reference whereby expressions denote themselves, rather than their customary denotation" (Abbott 2003:13), corresponding thus to what has been called the 'meaning of quotation marks' in the philosophical literature (Benbaji 2003).

(31) Sakha (BP2003, PotP_069)
*Attïal büt-en χolkuos **die-n** buol-ta*

cooperative end-PF.CVB collective.farm say-PF.CVB COP-PST.PTC.3SG *onton.*
then
'The cooperative ended and it then became a «collective farm».'

The metalinguistic function of non-canonical SAY is usually found in various naming constructions. It is obvious that this use of non-canonical SAY has developed out of its dissociative, quotative function: both quoting and the metalinguistic use "mention entities of the linguistic world instead of referring to phenomena in the object world" (Güldemann 2008: 399). Depending on the syntactic context of the SAY-marked expression, its lexical and syntactic nature, and the kind of reference it conveys, six types of metalinguistic non-canonical SAY can be distinguished.

A word or phrase followed by non-canonical SAY can be used as a predicative complement of a verb of calling or naming (**Nam_CallV**); most commonly, this structure applies to personal names (32), but other types of expressions also occur (33). In a few languages, this construction is not restricted to verbs of calling, but also appears with other types of verbs which can take predicative complements of the metalinguistic type (**Nam_DiffV**; (34)). The same structure – metalinguistically used word or phrase followed by non-canonical SAY – can furthermore be used to introduce naming expressions as main predicates, with or without a copula (**Nam_Pred**). This construction is almost exclusively found with personal names (ex. (35) – but see ex. (31)).

(32) Kolyma Yukaghir (Nikolaeva 2004: 41.1)
(...) *Jel'izar* ***mon-u-t*** *ńeː-ńə-ŋaː.*
Elizar say-0-IPF.CVB call-PROP-TR.3PL
'(...) they called him Elizar'.

(33) Udihe (Nikolaeva et al. 2003: p.128, 26.4)
(...) *uta-wa* *gegbi-si-ti* *cu'ai* *jaː-nie* ***gumu.***
that-ACC call-IPF-3PL bird.cherry flat.cake-3SG SAY.PTL
'(...) they call it bird-cherry flat cakes.'

(34) Western Ėven (DM2007, S.Golikov_101)
Kandiga *gerbe-w* (...) *Kaːn-Tiːgar* ***goːniken*** *dụk-kọːt-ta*

Khandyga name-ACC Kan-Tigar say.SIM.CVB write-GNR-NFUT(3PL)
'They write the word «Khandyga» as «Kan-Tigar» ...'

(35) North Mongolian (NMV 1974: p.30, 31.101; our glosses)
Matar Džögiin xaːni hüü , möŋgöŋ bodisat ***ge-dž*** *bae-na-w.*
M. Dz. Khan son silver B. say-IPF.CVB COP-PRS-1SG
'I am the son of Matar Dzögin Khan, Silver-Bodhisattva.'

In addition, non-canonical SAY is frequently used NP-internally in a kind of appositive construction (Nam_App): the head noun is preceded by another noun plus SAY, which denotes the name of the referent of the head noun (36).

(36) Dolgan (Eugénie Stapert field data 2008, Story_Vol_ANS_012)
Didipte ***die-n*** *üreχ iraːχ* (...)
Dudypta say-PF.CVB river far
'The river called Dudypta is far away (...)'

In North Mongolian and Eastern Evenki the form of non-canonical SAY used in Nam_App differs from that used in other SAY constructions: a habitual or imperfective participle (cf. (9) and (37) below), as opposed to the dominant, converbal form of non-canonical SAY (cf. Section 3.2). In Tuvan and Shor, too, participles are used with this function, but converbs of SAY predominate.

(37) Eastern Evenki (Romanova & Myreeva 1964: p. 161, 10.16; our glosses)
Tyvevul ***gune-riː*** *amiːkaːn-mi ulgučeːn-čeː-n*
Tyvevul say-IPF.PTC grandfather-1SG tell-PST-3SG
'This story was told by my grandfather (called) Tyvevul'

Words and phrases used metalinguistically can also be marked with non-canonical SAY if they function as independent semantic terms (Nam_Term), most often as arguments of a predicate (cf. (6) and (38)).

(38) Western Ėven (BP2010, BurID_traditions_095)

(...) *gọlọmọ* ***go:niken*** *įlụmụ-wčį-kan* *ọ:-p-tį*

yurt.type say.SIM.CVB yurt-SML-DIM make-MED-IPF.PTC

'(...) a «golomo» was made like an ilumu (a type of yurt).'

In a typologically less common construction, non-canonical SAY functions as a marker of metalinguistically used clauses and VP-like elements (**Nam_Cl/Rel**). The latter refer not to states of affairs, but to their own verbal content, and are thus a kind of pseudo-quote; at the same time, they are used as arguments of a higher predicate (39). This construction can occasionally transcend the narrow metalinguistic usage, so that non-canonical SAY appears simply as a marker of clausal or VP-like arguments (40), or, in a structure resembling relative clauses, as a marker of complex modifiers of noun phrases (41).[11]

(39) Sakha (BP2003 -Chir_081)

Dʒe, bu haχχa-li: iarij-di-lar ***die-n*** *buol-ar.*

well this Sakha-ADV be.sick-PST-PL say-PF.CVB COP-PRS.PTC

'That was (what was called) «they fell ill in a Sakha way».' (talking about appendicitis, which was believed to be a purely Sakha illness)

(40) Tuvan (TNS 1994, p. 316, 9.200; our glosses)

Meeŋ ko'dan-ym-ny olčal-ap al-yr ***de-p***

my encampment-1SG-ACC occupy-CNJ.CVB SBEN-PRS.PTC say-CNJ.CVB

čüve yndyg belen eves.

thing such easy NEG

'It is not so easy to occupy my encampment.'

(41) Kolyma Yukaghir (Nikolaeva 2004: 36.10)

Tude l'ə-gi maγil-gi (…) *jolo-γu la:γ-ə-t ta:tme:d'o:n*

his HST-3 coat-3 back-ADV.DIR to-0-ADV.ABL such

[11] The examples of the 'extended' type exemplified by (40) and (41) are very rare in our corpora and are obviously derived from the type exemplified by (39). For these reasons, we decided to treat them as instances of one single category in our statistical analysis, even if, strictly speaking, the former do not have the metalinguistic semantics characterising the latter.

moro-dəllə ***mon-u-t*** *dinna:q* *maɣil-ək.*
put.on-PF.CVB say-0-IPF.CVB indeed coat-FOC
'His coat (…) was a kind of coat that one puts on from the back.'

In terms of areal distribution, the following groupings are discernible. First, the languages of the northwestern (Enets) and southeastern fringes (Nivkh and Nanai) do not seem to use non-canonical SAY in metalinguistic constructions at all. In Udihe, geographically and genealogically close to Nanai, Nam_CallV is attested, but all other functions, including the otherwise highly prominent Nam_App, are absent. Second, on the other side of the scale, metalinguistic functions are so frequent in the Sakha life histories that they cover almost a third of all instances of SAY in the corpus.[12] Third, predicative usage types of non-canonical SAY – Nam_CallV, Nam_DiffV and Nam_Pred – are predominantly found in the northeast of the SAY area. Fourth, Nam_Term appears in the Turkic languages, where it might be a Siberian Turkic innovation, in North Tungusic, where it competes with an indigenous construction using *gerbe* 'name', and in Kolyma Yukaghir, where it is probably a contact-induced innovation (cf. Section 5.5). Finally, in the case of Nam_App, which is attested almost universally, distinctions can be made on the basis of the form used: in the south, N.Mongolian, E.Evenki, and to a certain extent Tuvan and Shor, use participles; other languages only use converbs in this function.

3.3.6. Non-canonical SAY in discourse-related functions

Non-canonical SAY often assumes functions related to communication management and the encoding of interpersonal or information-structural meanings (e.g. Chappell 2008: 49, Güldemann 2008: 411ff). Three such functions can be identified in the languages of Siberia: use as a discourse particle, use as an enumerative conjunction, and use as a topic marker.

The use of non-canonical SAY as a discourse particle (**Disc_Ptl**) is widely attested: forms of SAY appear in specific syntactic slots (sentence finally or at the right edge of the domain over which they have scope) and modify the proposition with various types of interactional, non-truth-conditional meanings. Disc_Ptl constructions can have

[12] Importantly, no such preponderance of metalinguistic SAY is apparent in the corpus of Sakha folklore (cf. Section 5.4 on the influence of genre).

various types of meaning: focal emphasis (42), different types of contrast (43), hedges (often in approximate quantification and after general extenders of the type 'etc.', 'and stuff', as in (44)), and others.[13] It is possible that some or most of these meanings are derived from the function of illocutionary reinforcement roughly corresponding to the English 'I'm TELLING you, don't do that again', i.e. from the basic dissociative function of non-canonical SAY (Güldemann 2008: 411). Most languages use dominant converbs to express discourse-related meanings; the exceptions are Tuvan and, partially, Buryat (see below), in which participles are employed.

(42) Dolgan (Eugénie Stapert field data 2009, Story_Syn_Holiday_MSA_32)
Öl-ör-büt-e ***die-n*** *du* *χas*
die-CAUS-PST.PTC-3SG say-PF.CVB Q how.much
'He did kill some (sc. reindeer), right?'

(43) Tuvan (Harrison 2005, 25.318)
(...) *čed-e* *maŋna-p* *kel-gen* *Boktu-Kiriš*
reach-CVB run-CNJ.CVB CLOC-PST(3SG) B.-K.
de-er-ge (...)
say-PRS.PTC-DAT
'[Hah! Before the guys had even reached the halfway point of the race-course, (…)], it was Boktu-Kiriš that came running up to that place.'

(44) Sakha (BP2002 Zhilinda_Lukinov_060)
Onton *bu* *anï-gï* *kurduk* *bu* *oɣuruot* *ah-ïn*
then this now-ADJR like this vegetable.patch food-ACC.POSS3SG
eŋin ***die-n-i*** *bil-bep-pit* *bukatïn.*
etc. say-PF.CVB-ACC know-PRS.PTC.NEG-1PL completely
'Then we didn't know these vegetables and such like nowadays at all.'

[13] Given this diversity of meanings, different types of Disc_Ptl should be treated separately, not subsumed under one category, as we do. However, the data we have often do not allow for clear ascription of meanings to the particle-like tokens of SAY found in the texts, and judgments of native speakers were available only for some of the languages in our sample. Therefore, we considered it safer to lump all subtypes together than to base our classification on speculations.

A special illocution-indicating type is found in Buryat, in which questions are often marked by the present participle *geeše*, augmented with the question morpheme *-b-,* which may agree in person and number with the subject:

(45) Buryat (BNS 2000: p. 70, 13.24; our glosses)
Ene jaa-ža baj-na ***ge-eše-b-ši*** (...)?
this do.what-IPF.CVB COP-PRS say-PRS.PTC-Q-2SG
'What are you doing (...)?'

A function practically restricted to Sakha (with only rare attestations in Tuvan and Buryat; cf. Bertagaev & Cydendambaev 1962: 223ff), is the use of non-canonical SAY as a conjunction in listing constructions (**ENUM**). Even though this is not a discourse-related use in the narrower sense of the word, we describe it together with Disc_Ptl because of its semantic and formal closeness to the type of discourse particle which expresses hedges in approximate quantification and after general extenders ('etc.'; ex. (44)); it is possible that it has originated in this latter function.

(46) Sakha (BP2002, Efmy_401)
Bu oɣo-ɣo usturuːs ***die-n,*** *tuoχ untuː ullar-ar kilej*
this child-DAT plane say-PF.CVB what unty sole-PRS.PTC glue
die-n, *hap* ***die-n,*** *aniː erehiːne tiː hakkaːs-tiː-lar.*
say-PF.CVB thread say-PF.CVB now rubber boat order-PST-PL
'This child wants a plane, and glue for soling unty (fur boots), and thread, and they ordered a rubber boat.'

As a topic marker (**TOP**), non-canonical SAY is attached directly to the topic expression, which is in turn moved to the left edge of the clause. The contexts in which this type of topic marking is found are usually those of different types of topic shift – contrast, resumption of an old topic, establishment of a new one, etc. The topic-marking function of non-canonical SAY is probably related to its metalinguistic use, and thus ultimately to its basic, dissociative function: an entity is first named (which is marked by non-canonical SAY), and then predicated about. This function is usually encoded by dominant SS converbs (N.Mongolian, Buryat, Sakha, Dolgan), but

conditional converbs (Shor) and participles (Tuvan) are found in Southern Siberian Turkic.

(47) Tuvan (Harrison 2005: 12.114-5)
Xamнk küdee-ler ***de-er-ge*** *čıg-l-ıp* *kel-gen (…)*
all suitor-PL say-PRS.PTC-DAT gather-PASS-CVB come-PST(3SG)
'As for all those suitors, they have already gathered (…).'

3.3.7. Non-canonical SAY as a marker of the standard of comparison

In the Western Ėven life history corpus, but nowhere else in our sample, non-canonical SAY can be used to mark the standard of comparison in comparative structures (**COMPR**). This construction is very rare, but is confirmed by our informants. The relationship of this function of SAY to other functions is not entirely clear.

(48) Western Ėven (BP2008, Stepanova_ZA_svatovstvo_30)
(...) *aja:w-rị-n* *asatka-l-dụk* ***go:-mi*** *er* *ńarị-ka-r*
love-PST-3SG girl-PL-ABL say-COND.CVB this man-DIM-PL
kụŋa-l-bụ.
child-PL-ACC
'(She) loved boys more than girls.'

3.3.8. Lexicalised and idiomatic uses of non-canonical SAY

In a number of Siberian languages, the combination of non-canonical SAY and question word (particle or adverb) is lexicalised as a causal conjunction (**Lex_Caus**), with a meaning corresponding to English 'because' (Pal'mbax 1955: 182, Aydemir 2009: 122; (49)). All the forms roughly mean "if one were to say why". Sakha, Tuvan, and Buryat use the oblique case of a participle (Sakha *dieteχχe*, Tuvan *deerge*, Buryat *gexede*), N.Mongolian and W.Ėven an impersonal conditional converb (N.Mong. *gebesü,* W.Ėven *go:mi*). This is preceded by a question word (Sakha *toɣo,* Tuvan *čüge*, N.Mong. *jaγan*, Buryat *juum*, W.Ėven *ịamị*). The pattern seems to be a Turkic-Mongolian phenomenon (cf. also Weiers 2003: 262 on Mogghol), Dolgan being the only Turkic language in the sample lacking it. As for W.Ėven, this is probably a calque from Sakha, judging from dialectal evidence: while *toɣo dieteχχe* is present in all

Sakha dialects, *ɟamɨ goːmi* is found only in one of the subdialects of W.Ėven (the Lamunkhin dialect), which has a prolonged and intensive history of contact with Sakha (Pakendorf 2009).

Apart from the 'because'-lexicalisation of SAY, which occurs most frequently, other sporadic instances of lexical conventionalisations are found in the South Siberian languages (**Lex_Other**). These often occur in combination with adverbial expressions, such as Tuvan *araj dep* (barely say-CNJ.CVB) 'hardly', Buryat *xajšan geed* (whither say-PF.CVB) 'how, in which way' and Shor *noo deen* (what say-PST.PTC) 'of what quality, how', but also on their own, like N.Mongolian *ge-tel(e)* (say-TERM.CVB) 'while, meanwhile' (50) and Shor *te-ze* (say-COND.CVB) 'however; or'. Even though our data in this domain are far from complete, we can tentatively conclude that a limited number of lexicalisation patterns (adverb + converbal SAY, converbal SAY as conjunction) are very productive in South Siberia. This justifies including these different lexical units in one category.

(49) Buryat (BNS 2000: p.58, 8.4; our glosses)
Ügy, ende edi-že bolo-xo-güj, ***juum ge-xe-de***
no here eat-IPF.CVB become-FUT.PTC-NEG Q.PTL say-FUT.PTC-DAT
ene xargy-gaar eldebyn amitan jaba-na
this road-INST different animal go-PRS
'No, one shouldn't eat here, because various animals go along this road.'

(50) North Mongolian (NMV 1974: p.29, 31.44; our glosses)
Ge-tl *tsuk jaw-dž-esŋ xüŋ neg ünüg šarxaduula-dž.*
say-TERM.CVB together go-IPF.CVB-?? man one fox wound-PST(3)
'While they were riding there together, a man wounded a fox.'

Yet another type of semi-lexicalised structure is parenthetically used expressions with non-canonical SAY (**Lex_Par**). These usually contain a demonstrative adverb (Dolgan *ol dien* [that say.SEQ.CVB], W.Ėven *ečin goː-niken* [so say-SIM.CVB], both: 'so, so to say'), the name of a language (W.Ėven *ńuːčidi-t goː-mi* [Russian-INST say-COND.CVB] 'in Russian'), or an interrogative (W.Ėven *ɟak goː-niken* [what say-SIM.CVB] 'whatchamacallit'). These idiomatic expressions are used as a side

comment on the main predications and are syntactically independent insertions into the main clause.

(51) Dolgan (Eugénie Stapert field data 2008, Story_Vol_LKS_155)
Ol ***die-n,*** *d'e honon anï-ga dieri Dudinskij oruojoŋ-ŋa*
that say-PF.CVB PTL that.way now-DAT till Dudinka.ADJ district-DAT
olor-obut, barikaːm-mït
live-PRS.1PL all-1PL
'So, that's how we all until now live in the Dudinka district.'

3.3.9. SAY as auxiliary/light verb

The use of SAY as an auxiliary or a light verb is formally different from all functions of non-canonical SAY dealt with in this paper. While all other types of non-canonical SAY are restricted to non-finite verb forms, this type has full-fledged inflection – which is only natural, given that the main function of auxiliaries/light verbs is to carry TAM features. Semantically, however, it falls under our definition of non-canonicity, since the meanings of the auxiliary/light verb constructions are not compositionally derivable from the meaning of SAY.

Used as an auxiliary verb (**AUX**), non-canonical SAY functions as the head of a phrase consisting of the auxiliary and a non-finite form of the lexical verb. The most important periphrastic structure in which the SAY-based auxiliary plays a role is Intentional, in which SAY is complemented with a future participle (cf. Mongush 1987:88 on Tuvan, Čeremisov 1973: 174, Skribnik 1987: 43, 45 on Buryat). This periphrastic structure probably stems from a biclausal construction with direct speech (roughly: 'I say/think I will...') and is thus clearly related to the basic dissociative function of SAY. It is restricted to the Mongolic languages and Tuvan in South Siberia.

(52) North Mongolian (NMV 1974: 4, 24.19; our glosses)
(...) *xüːg öːriːn ide-xe* ***ge-dž,*** *maŋgas.*
boy self eat-FUT.PTC say-PST(3) mangas
'(...) she wanted to eat up the boy herself, the mangas-witch'

SAY can also occur as a light verb (Light_V) in conjunction with ideophones (Čeremisov 1973: 174, Pal'mbax 1955: 181). In this structure, widely attested in Africa (e.g. Creissels 2001:80ff, Güldemann 2002: 260ff), SAY functions as a proper light verb, i.e. it combines with non-inflected lexemes and serves as the carrier of TAM features when they are used as main or embedded predicates. It is obvious that Light_V is closely related to the dissociative complex of functions, more specifically to metalinguistic and, ultimately, quotative functions of SAY (cf. also Güldemann 2008: 280ff). However, as noted above, the syntactic behaviour of non-canonical SAY as light verb is different from that of other dissociative functions: it has a full-fledged paradigm and broad syntactic distribution, while other functions have petrified forms and are limited to one or two syntactic contexts. Similar to AUX, Light_V is found only in Mongolic and Tuvan.

(53) Buryat (BNS 2000: 38, 1.147; our glosses)

Tiixeden	*baabgajn*	*teb-teb*	***ge-xyn***	*duula-ad.*
then	bear	IDEO:chew	say-FUT.PTC.GEN	hear-PF.CVB

'Then the bear heard (the fox) chewing.'

3.3.10. Unclear cases

In almost all of the languages in the sample, there were instances of non-canonical SAY which we were not able to analyse in terms of function; their numbers range from 18 (2.7%) in the Sakha life history corpus and 15 (1.6%) in the W.Ėven life history corpus to 2 (0.4%) in the E.Evenki corpus and 1 (0.7%) in Enets.

3.3.11. Functions of non-canonical SAY: A summary

Most of the numerous functions which non-canonical SAY can have in Siberia are derived from the basic function of dissociation. In the preceding sections, we have repeatedly indicated the connections between different functional domains, some of which must, for lack of historical evidence and due to the nature of our material, remain hypothetical. The graphic summary in Figure 4 shows the (assumed) relationships among functions: The fields of direct verbal quotation and internal awareness make up the central complex of dissociative semantics (shaded gray); this central functional domain is in various ways connected to other functional domains.

Note that the schema of relations we propose does not purport to represent a diachronic picture or grammaticalisation path, as we believe that, with the data at hand, any such attempt could only be highly speculative. It is rather intended to reflect formal and semantic similarities between functions or groups of functions. These similarities can be so great as to make the classification of particular tokens of non-canonical SAY difficult, as illustrated repeatedly in the preceding sections – a situation which, to emphasise the point again, is characteristic for the domain of non-canonical SAY, which is to be conceived of as a continuum rather than as a set of discrete functions.

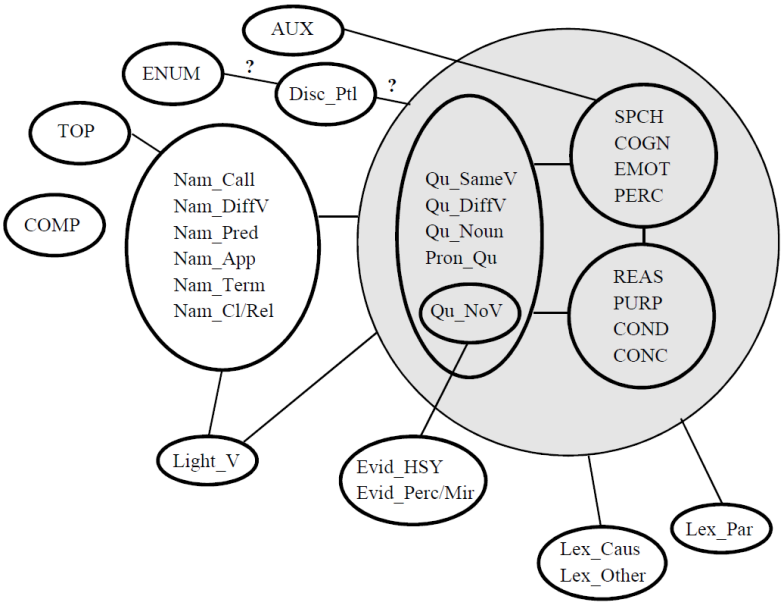

Figure 4: Schema of relationships among functions of non-canonical SAY in the languages of Siberia

AUX – auxiliary; COGN – complementiser with verbs of cognition; COMP – standard of comparison; CONC – marker of concessive clauses; COND – marker of conditional clauses; DISC_PTL – discourse particle; EMOT – complementiser with verbs of emotion; ENUM – enumerative particle; EVID_HSY – hearsay evidential; EVID_PERC/MIR – perceptual/mirative evidential; LEX_CAUSE – lexialised 'because'; LEX_OTHER – lexicalisations other than 'because'; LEX_PAR – lexicalised parentheticals; LIGHT_V – light verb; NAM_APP – metalinguistic marker in appositive structures; NAM_CALL – metalinguistic marker with verbs of calling; NAM_CL/REL – metalinguistic clause marker; NAM_DIFFV – metalinguistic marker with verbs other than verbs of calling; NAM_PRED – metalinguistic marker with nominal predicates; NAM_TERM – metalingustic marker with semantic terms; PERC – complementiser with verbs of perception; PRON_QU – pronominal quote; PURP – marker of purpose clauses; REAS – marker of reason clauses; QU_DIFFV – quote marker with a verb other than SAY; QU_NOV – quote marker without a governing verb/noun; QU_NOUN – quote marker dependent on a noun; QU_SAMEV – quote marker with SAY; SPCH – complementiser with verbs of speech; TOP – topic

4. Analysis of the data and emergent patterns

As has become clear from the description of the frequency, forms and functions in which non-canonical SAY occurs in the languages of Siberia, this is a very heterogeneous domain. While for some language families the presence or absence of non-canonical SAY can be assumed to be an inherited feature (Mongolic and Turkic vs. Yeniseic, Uralic and Chukotko-Kamchatkan, respectively), for others, especially the Tungusic and Yukaghir languages, the origin of this phenomenon is far less clear. Furthermore, even within language families with a fairly clear pattern of inheritance we find outliers that do not follow the expected genealogical pattern, i.e. Enets within the Uralic language family and Dolgan within Turkic. Specifically, it is quite noteworthy that Dolgan, which is so closely related to Sakha as to be considered a dialect in some classifications, makes far less use of non-canonical SAY than its sister does in both text genres investigated (cf. Table 4 and Appendix 2). It is thus to be expected that inheritance, convergence and independent language-internal developments have all played a role in producing the variegated pattern we see among the languages of Siberia today. Before we try to identify some of these developments in Section 5, we here present the results of exploratory analysis performed to discern patterns in the data.

Using the programme Statistica (ver.10, StatSoft Inc.), we performed a Correspondence Analysis (CA) on the absolute number of tokens attested for each function in our corpora (including canonical uses of SAY but excluding unclear tokens). This analysis is similar to Multidimensional Scaling and Principal Components Analysis in that it extracts the most important components of information from a multidimensional space of variables and lets one visualize these in graphical space. It has the added advantage of allowing one not only to analyse the placement of variables – in our study, the individual languages – with respect to each other, but also enabling one to see which factors – particular functions of non-canonical SAY – are driving the relative positions of the variables. The full two-dimensional plot, which covers approximately 57% of the variation in the data, is given in Figure 5; Figure 6 provides a close-up of the central area of the plot.

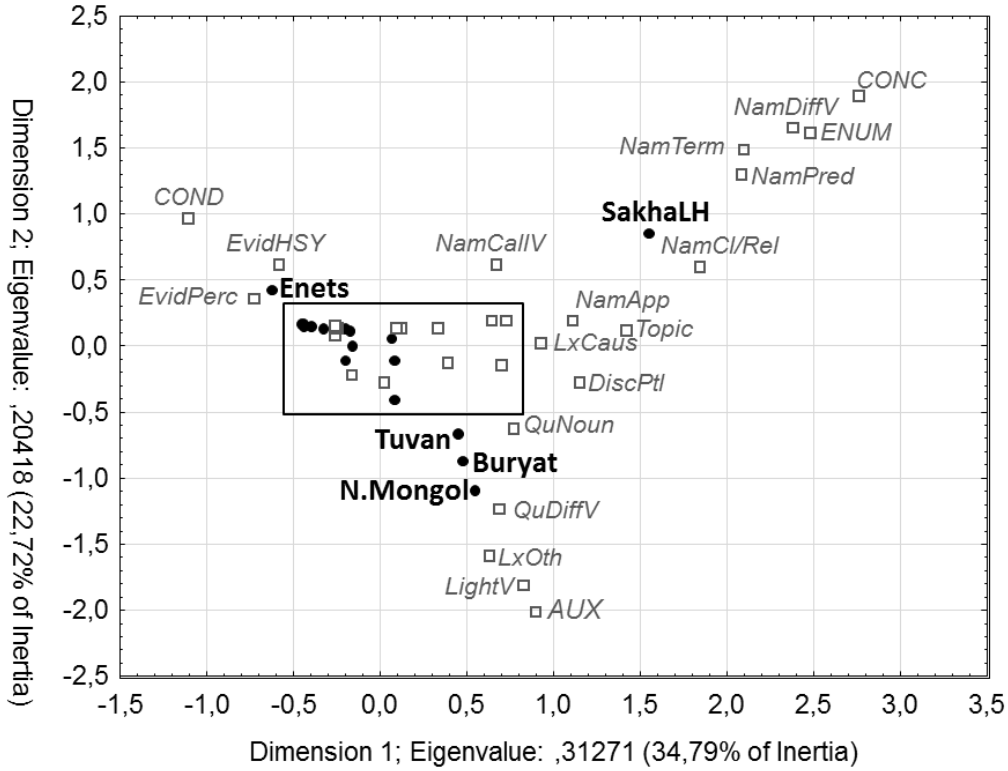

Figure 5: Two-dimensional CA plot (complete plot)

LH = life histories;

AUX – auxiliary; CONC – marker of concessive clauses; COND – marker of conditional clauses; DISCPTL – discourse particle; ENUM – enumerative particle; EVIDHSY – hearsay evidential; EVIDPERC – perceptual evidential; LXCAUS – lexialised 'because'; LXOTH – lexicalisations other than 'because'; LIGHTV – light verb; NAMAPP – metalinguistic marker in appositive structures; NAMCALL – metalinguistic marker with verbs of calling; NAMCL/REL – metalinguistic clause marker; NAMDIFFV – metalinguistic marker with verbs other than verbs of calling; NAMPRED – metalinguistic marker with nominal predicates; NAMTERM – metalingustic marker with semantic terms; QUDIFFV – quote marker with a verb other than SAY; QUNOUN – quote marker dependent on a noun; TOP – topic

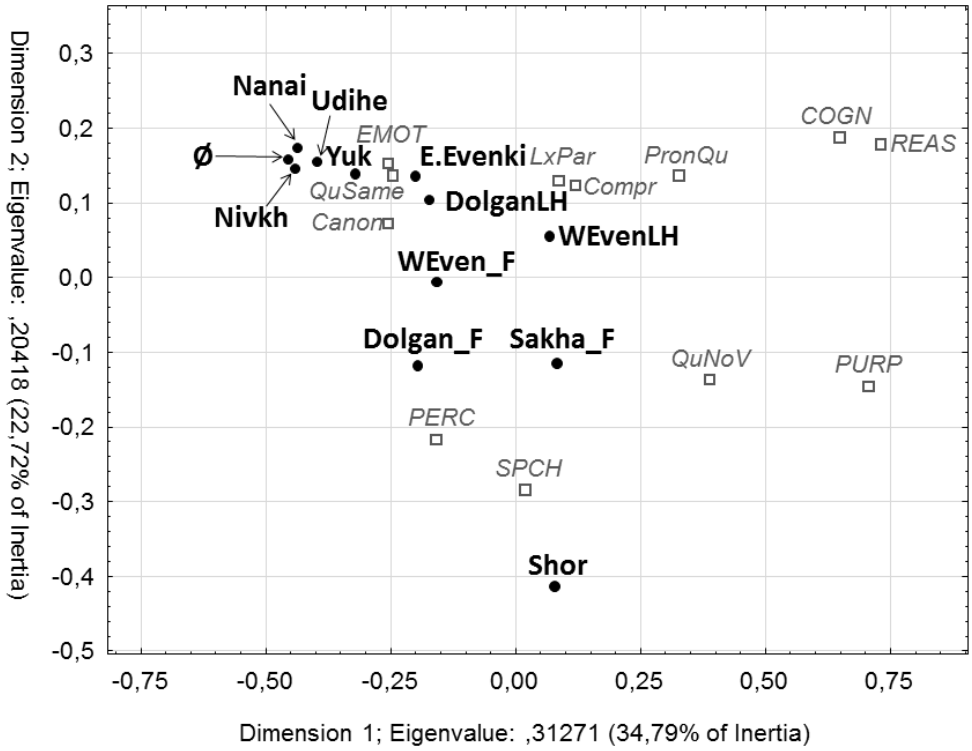

Figure 6: Two-dimensional CA plot (zoom-in of the central area)

F = folklore; LH = life histories; Yuk = Kolyma Yukaghir; Ø = languages lacking non-canonical SAY

CANON – canonical SAY; COGN – complementiser with verbs of cognition; COMPR – standard of comparison; EMOT – complementiser with verbs of emotion; LXPAR – lexicalised parentheticals; PERC – complementiser with verbs of perception; PRONQU – pronominal quote; PURP – marker of purpose clauses; REAS – marker of reason clauses; QUNOV – quote marker without a governing verb/noun; QUSAME – quote marker with SAY; SPCH – complementiser with verbs of speech

What this exploratory analysis shows is that there are clear subgroups of Siberian languages with regards to the use of non-canonical SAY. The plot distinguishes several outliers from a core group of languages, which is divided into a southeastern (South Tungusic, Nivkh) and a northern subgroup (E.Evenki, W.Ėven, Dolgan and the Sakha folklore corpus). While the southeastern subgroup is clearly set off by the predominance of evidential functions and the scarcity or absence of some frequent types of non-canonical SAY, such as quote, complement or adjunct marking, the

affinities between the languages of the northern subgroup are less clear-cut, providing some indication that there are complex factors at play here. Outside the core group, the South Siberian languages Tuvan, Buryat, and North Mongolian are united by very high frequencies of obligatory Qu_DiffV as well as by the use of non-canonical SAY as light verbs, auxiliaries, and in idiomatic lexicalized expressions. This clustering arguably reflects parallel inheritance in conjunction with convergent development due to long-standing close contact, as will be further discussed below. The Sakha life histories and Enets are also clearly separated from the rest, Enets due to the idiosyncratic use of SAY as conditional, the frequency of the evidential function and the lack of practically all other functions, and the Sakha life histories due to the high frequency of various metalinguistic uses of SAY and the presence of the enumerative and concessive functions.

Even though the picture might seem clear at first blush, there are intricacies which demand further scrutiny. For example, Yukaghir appears between the southeastern group and its geographical neighbours W.Even and Sakha, while Shor, which is spoken in South Siberia, is in a transitional position between the core group and its geographic neigbours Tuvan, Mongolic, and Buryat. Most of the groups revealed by the CA correspond to areal and/or genetic groupings, but the exact nature of the connections within and across groups is not immediately obvious. In addition, while genre seems to be the major determinant of the position of Sakha in the plot, its effect on the W.Even and Dolgan data seems comparatively minor. In the following section we will try to disentangle some of these complexities in order to arrive at a better understanding of the processes involved in the development of the Siberian SAY-area.

5. Historical and areal interpretation: the spread of SAY in Siberia

As indicated in the introduction, non-canonical SAY is ubiquitous in the languages of the world. It is also an inherited feature of at least some languages of Siberia, which furthermore have a long history of language contact. The presence of non-canonical SAY in this region can thus be due to three factors: independent internal developments, inheritance from a common ancestor language, or contact influence. In what follows, we shall attempt to unravel these three factors on the basis of the data presented in Sections 3 and 4. It should be noted at the outset, however, that these factors are not mutually exclusive: (a) different types of non-canonical SAY in one language could

have developed due to different factors, (b) inherited functions of non-canonical SAY may have been expanded or abandoned under contact influence, and (c) independent developments could be interacting with both inherited predispositions and contact pressures.

5.1. Northwest

The independent development of the attested non-canonical functions of SAY in Enets is easy to demonstrate: being a Uralic language, it cannot have inherited non-canonical SAY, since this feature is largely absent in that family (see Section 3.1). Furthermore, it stands out among the languages in our sample by lacking the most widespread functions (quote and complement marking functions, as well as REAS and PURP) while making use of idiosyncratic forms of SAY in functions not widely attested in the languages of Siberia (Evid_HSY and especially as a marker of conditional clauses). Since no other SAY-language of Siberia with which Enets is or was in contact (currently Dolgan, in earlier times Tuvan and its relatives, Helimski 2003) uses these forms or has any of these functions, independent development is the most plausible explanation. This again attests to the inherent tendency of generic verbs of speech to take on functions not related to their primary lexical semantics.

5.2. Southeast

A small grouping that emerges from the analysis is that of the South Tungusic languages Udihe and Nanai and the isolate Nivkh. Two features unite these languages: the high frequency of evidential uses of SAY and the paucity or complete lack of other functions. Within South Tungusic, an additional shared feature is the use of non-canonical SAY as a marker of perceptual/mirative evidentiality, a function not found elsewhere in Siberia. Whether this arose in a common ancestor of the two languages and was inherited by them, or whether it arose in one language and was transmitted to the other by contact is difficult to judge. However, given the divergent forms used by these two languages (participle in Udihe vs. converb-derived clitic in Nanai), a shared ancestral source appears somewhat unlikely, as discussed below in more detail for the Tungusic language family as a whole.

As to Nivkh, without known linguistic relatives it is impossible to establish the presence or absence of non-canonical SAY in Proto-Nivkh. The synchronic evidence

is contradictory. On the one hand, non-canonical SAY in Nivkh has idiosyncratic forms (a bound morpheme occurring only as a quote marker and an evidential particle), so that contact-induced developments appear improbable. On the other hand, the sharing of features with its South Tungusic neighbours renders at least some contact influence possible. Thus, while contact-induced developments are plausible for the southeastern area, they cannot be determined with any certainty.

5.3. South

One of the most secure micro-areas within the larger Siberian area investigated here is represented by the South Siberian languages Buryat, N.Mongolian, and Tuvan. These are united by similar frequencies and functions of non-canonical SAY, which are at least partly inherited (cf. Section 2.3). Among the functions that both the Turkic and the Mongolic languages seem to have inherited are Quote_DiffV, complementizer with cognition verbs, and as a marker of names in apposition and with verbs of calling (cf. Tables 2 and 3). As to the discourse-related functions of SAY, which are very salient in the modern-day Turkic and Mongolic languages of Siberia, their historical status is difficult to assess in the absence of any mention in the largely traditional descriptions of the languages belonging to these language families. We are aware only of a couple of scattered attestations across Turkic and Mongolic (Turkmen – Clark 1998: 455, Turfan Uyghur – Yakup 2005: 148-9; Classical Mongolian – Lessing 1960: 372, Kalmyk – Muniev 1977: 141-5), which at best show that some of the discourse functions might be ancient. The sharing of other features between (some of) the Turkic and Mongolic languages included in our sample, however, might well be due to convergent developments through contact, with the contact influence going in both directions. For instance, while purpose and reason adjuncts are certainly inherited in the Turkic languages (cf. Section 2.3), these functions are not attested in the early Mongolic monuments, nor do they appear in Dagur (Martin 1961: 150) and the Shirongolic branch (Bao'an – Fried 2010: 294ff, Mangghuer – Slater 2003: 309), which are both outside the sphere of Turkic influence. However, they are common in the Central branch (in addition to Buryat and Northern Mongolian, in Kalmyk – Muniev 1977: 142, and Standard Khalkha – Slater 2003: 309), making development under contact influence from Turkic possible for at least some of these languages. Conversely, the only Siberian language family in which the auxiliary and light verb

uses of SAY appear to be ancient is Mongolic (cf. Table 3 above), while there is no evidence for an old origin of these functions in the Turkic language family. Therefore, their presence in Tuvan is most likely due to contact influence from the surrounding Mongolic languages, with which Tuvan has been in intense contact (see Khabtagaeva 2006 on Mongolic-Tuvan linguistic relationships).

The position of the Turkic language Shor is indicative in this respect. Shor lacks those defining features of the southern group that stem from Mongolic, i.e. the auxiliary and light verb uses, which presumably reflects the lower intensity of Mongolic influence in Altai Turkic, to which Shor belongs (cf. Schönig 2003: 412ff.). On the other hand, it displays all the typical inherited Turkic features enumerated above. The corollary of this is the intermediate position of Shor between the South Siberian group, characterised by a mixture of features inherited from Mongolic and Turkic, and the northeastern group, the history of which seems to be much more complex, including both Turkic influence and other types of developments. Before turning to the description of this group, a discussion of the influence of genre on our classification is necessary, since it is among the languages of the northeast that genre differences influence the position of some languages in our areal model.

5.4 The problem of genre and the position of Sakha

As described in Section 2.1, the corpus on which this study is based consists of narrative texts, both transcribed field data and published stories. The bulk of the corpora consists of folklore tales, a genre with a relatively stereotypical discourse structure (linear temporal frame, long dialogues, repetitions, etc.) and a restricted repertoire of participating entities and actions. We assume that the inherent characteristics of this genre render the corpora sufficiently comparable, although we cannot exclude an uncontrolled-for impact of the editing process on the published texts (see Section 2.1). In the case of Sakha, Dolgan and W.Ėven, however, the issue of genre is graver than elsewhere: our field data from Sakha and Dolgan consist almost exclusively of life histories, and in W.Ėven, life histories are roughly five times more frequent than folklore texts. Although life histories are also narrative in nature, they are structurally much more diverse than folklore stories and have no pre-established set of participants and situations; furthermore, they lack the dialogue characteristic of folklore texts. As described in Section 2.1, we created two separate corpora for each

of these three languages in order to control for the influence of genre on our results (cf. Appendix 1).

There are differences with respect to the overall frequency of non-canonical SAY between the two genres in Sakha and W.Ėven (cf. Table 4), with far more non-canonical SAY being found in the life histories than in the folklore texts. The frequency of some functions appears to depend on genre, too (Appendix 2). Thus, the quotative use (Qu_DiffV) of non-canonical SAY is more frequent in the folklore data, doubtlessly due to the higher incidence of dialogues in this text type; conversely, discourse-related functions are less frequent or absent in the folklore data, probably in consequence of the non-interactive nature of this genre and/or editing practices.

This minor variation notwithstanding, genre does not have a notable influence on the position of the language in the CA plot for W.Ėven and Dolgan (Figure 6). In contrast, the position of Sakha is fully determined by genre: while the folklore corpus of Sakha clusters together with its closest relative, Dolgan, and its closest neighbours, W.Ėven and E.Evenki, the corpus of Sakha life histories is separated from all the languages included in our study (Figures 5 and 6). Some of the differences between the Sakha folklore and the Sakha life histories are due to the characteristic elements of folklore texts, such as a high amount of direct and reported speech; this probably makes the Sakha folklore corpus more similar to the other northeastern languages. However, the same genre-specific elements are observable in Dolgan and W.Ėven, too, but no such profound differences between the two text types surface: for both languages, both folklore and life history corpora are firmly rooted in the northeastern area. This makes it clear that the closeness of the Sakha folklore data to the languages of the northeast is not an artefact of the genre, but is indicative of the areal connections of Sakha, and it is the Sakha life histories that are really outstanding. The reasons for the separation of the Sakha life histories, as shown in the CA plot, are an extremely high incidence of various metalinguistic functions of non-canonical SAY and the presence of the concessive function in this corpus as well as, though to a lesser degree, the presence of the topic-marking and enumerative function. The major factor triggering the extremely high frequency of metalinguistic functions in the life histories (as opposed to the medium frequency found in the folklore tales) is the type of interaction, in which the speakers introduce both themselves and a multitude of new referents and concepts which they assume are unknown to the interviewer; a similar, though much less pronounced, difference in frequency of these functions between the

folklore and life history data can be observed in W.Ėven and Dolgan. The frequency of the topic-introducing function in oral life histories can be explained along the same lines, since it is commonly new topics that are introduced with non-canonical SAY; the enumerative function, too, might be related to the discourse type. The fact that only life histories, but not folklore tales, contain tokens of the the concessive function might be due to the innovative nature of the concessive construction, which hasn't yet found its way into traditional narratives;[14] however, this claim is difficult to substantiate.

This short analysis shows that the influence of genre should not be underestimated in a corpus-based study like ours, since it is hard to predict and difficult to control. While the impact of the text type on non-canonical SAY in Sakha is profound, it seems to have only minor effects in Dolgan and W.Ėven. To ensure inter-corpora comparability, the discussion of the northeastern group in the following section will be based mainly on the folklore corpora of these three languages, resorting to the data from the life histories only to illustrate historical connections which might otherwise go unnoticed.

5.5. Languages of the northeast and Dolgan

The most complex grouping that emerges from the Correspondence Analysis is that of the northeastern languages Sakha, E.Evenki, W.Ėven, and, marginally, K.Yukaghir, as well as Dolgan from the northwest. The history of this micro-area is far more difficult to unravel than that of the South Siberian languages. First of all, the diachrony of non-canonical SAY in Tungusic and Yukaghir cannot be established with the same certainty as in the Turkic and Mongolic languages, making it an arduous task to assess the probabilities of inheritance vs. independent developments vs. contact-induced change. The difficulty with reconstructing ancestral states in these languages arises from the lack of ancient sources: earliest attestations (word lists collected by non-specialists) go back no further than a few centuries, with textual data mostly going back no further than the late 19th century. Furthermore, with only two languages remaining in the Yukaghir family (Kolyma Yukaghir, which is included in our sample, and Tundra Yukaghir), both of which are spoken in a contiguous area with very similar

[14] The folklore texts analysed here were recorded mainly in the 1930s and 1940s, with two from the 19th century and three recorded in the 1970s and 1980.

contact situations, we lack good outgroups for comparison. As to the Tungusic languages, the distribution of non-canonical SAY is so patchy that it is hard to come to reliable conclusions concerning the ancestry of the feature. Secondly, E.Evenki, W.Ėven and Yukaghir have all been in documented contact with Sakha, the sociolinguistically dominant language of the area, for more than a century (Wurm 1996: 976); as a Turkic language, Sakha has clearly inherited several functions of non-canonical SAY. This raises the possibility that some of the functions of non-canonical SAY attested in the northeastern languages might have arisen under influence of Sakha contact, but without historical data this is hard to prove. Nevertheless, we will attempt to elucidate at least some of the potential factors at play in the establishment of this micro-area.

As mentioned above, the question concerning the origin of non-canonical SAY in the Tungusic language family poses a serious problem that we are unable to solve here. On the one hand, this feature is attested in all branches of the language family (North Tungusic, South Tungusic, and Manchu – see Gorelova 2002: 273-276, 353-354, 526-528 for the latter), leading to the conclusion that it might be an ancient feature. On the other hand, there are serious issues with this scenario: first of all, the functions of non-canonical SAY attested in the different Tungusic lects are strikingly different, as demonstrated by the results of the Correspondence Analysis, where the South Tungusic languages Nanai and Udihe are grouped separately from the North Tungusic lects W.Ėven and E.Evenki. Secondly, it is noteworthy that in the North Tungusic branch non-canonical SAY is absent with the exception of those lects that are in documented close contact with Sakha: E.Evenki (cf. Vasilevič 1948: 253-254, 301, 326; Romanova & Myreeva 1962, 1964) and W.Ėven (cf. Tugolukov 1997, Malchukov 2006). The western Evenki dialects and the eastern Ėven dialects, which are not in contact with Sakha, lack this feature entirely, as does the related North Tungusic language Negidal. This might be an indication that the development of non-canonical SAY in E.Evenki and W.Ėven is due rather to contact influence from Sakha than to inheritance, a supposition strengthened by the fact that the North Tungusic lect with the highest amount and the greatest diversity of functions of non-canonical SAY is the Lamunkhin dialect of Ėven (subsumed under W.Ėven together with the Tompo dialect in this study), which has undergone considerable contact-induced changes under Sakha influence (Pakendorf 2009). Furthermore, Manchu, too, is known to have undergone extensive contact-induced changes under the influence of Sinitic and

Mongolic languages (Gorelova 2002: 18-38), which in turn are well known to exhibit non-canonical SAY. Some of the functions attested in Manchu, such as light verb with ideophones and marker of conditional clauses, are not attested in the other Tungusic languages, but are attested in Mongolic and some Sinitic languages (cf. the discussion of the Mongolic languages above and Chappell 2008); similarly, the fact that Manchu occasionally employs a finite form of SAY in non-canonical functions (Gorelova 2002: 526) aligns it more with the Mongolic than the Tungusic languages. This raises the question whether at least some, if not all, of the functions attested in Manchu might not be due to contact influence rather than inheritance from a common Tungusic ancestor, further weakening the case for an ancient Proto-Tungusic origin of this feature. Third, a further difficulty with the scenario of an inherited origin of non-canonical SAY in the Tungusic languages is the variation in form: notwithstanding the fact that there is a shared pool of forms from which the Tungusic languages draw their non-canonical SAY (the simultaneous converb in *-*mi:*/-*mai*, the innovated North Tungusic converb in *-*nA*, and participal forms in *-*RI*), the South Tungusic languages Udihe and Nanai stand out in using a participle and particle (Udihe) and a clitic (Nanai; cf. Table 5) rather than converbs. Although the Nanai clitic arguably developed out of a converbal form of SAY, the fact of its cliticization and the lack of converbs in Udihe point to a separate development of non-canonical SAY in the North and South Tungusic branches, which also emerges from the analysis presented in Section 4 above. Thus, it is at this point impossible to identify whether non-canonical SAY in the Tungusic languages arose as a result of independent internal developments, contact pressure, or shared inheritance; however, the last possibility appears least likely, as outlined above.

Non-canonical SAY exists in both Tundra Yukaghir[15] and Kolyma Yukaghir (Proto-Yukaghir simultaneous converb in *-δ, Nikolaeva 2006: 83, Tundra Yukaghir *monur*, Kolyma Yukaghir *monut*), and many of the functions are comparable, such as quote, complement and adjunct clause marking. However, non-canonical *monur/monut* is not found in the earliest recorded texts (Jochel'son 1900), and, as mentioned, the contact situations of Tundra Yukaghir and Kolyma Yukaghir were so similar (with strong influences from Ėven and Sakha), that it cannot be ascertained to what extent non-canonical SAY in both languages is common heritage and to what

[15] Claims on Tundra Yukaghir are based on Dejan Matić's field data and the texts in Maslova (2001) and Kurilov (2005).

extent it is a product of parallel contact-induced developments. The intermediate position of K.Yukaghir between its geographic neighbours and the small cluster of southeastern languages plus those lacking non-canonical SAY entirely (cf. Figure 6) can be explained by the fact that a large variety of the functions characteristic of the northeastern languages (quote marker, complement marker, adjunct marker, metalinguistic functions) are attested in the K.Yukaghir corpus, but in very low frequencies (cf. Appendix 2). It is these low frequencies that locate K.Yukaghir closer to the languages lacking non-canonical SAY than to its geographic neighbours, which exhibit high frequencies of these functions.

However, while it is difficult to disentangle possible inheritance from contact influence for the North Tungusic languages and Yukaghir, metalinguistic functions of SAY represent a rare case in which areal influence can be ascertained with high probability. In contrast to K.Yukaghir, there are no instances of metalinguistic non-canonical SAY in T.Yukaghir. Instead, T.Yukaghir makes use of participles and converbs of the copula *ŋoːl-*, which is a feature not found in any of the neighbouring languages and is thus most probably not contact-induced. In contrast, the use of SAY in metalinguistic contexts in K.Yukaghir is a distinct areal feature, found in Sakha as well as in the North Tungusic lects. It is thus probable that metalinguistic non-canonical SAY is an innovation in K.Yukaghir, probably induced through contact with neighbouring languages. North Tungusic languages are the less likely source: metalinguistic functions are attested only very sporadically in Tungusic (cf. Appendix 2), with no mention in the grammars of Manchu (Avrorin 2000, Gorelova 2002), and are much more often expressed with what appears to be a genuinely Ėven construction in W.Ėven (the noun *gerbe* 'name') than with non-canonical SAY. It is therefore probable that metalinguistic uses of SAY are an innovation in E.Evenki and W.Ėven, making contact-induced developments under Sakha influence plausible for both the North Tungusic dialects and K.Yukaghir. This assumption is all the stronger since Sakha inherited at least some of its metalinguistic uses of SAY (cf. Table 2).

With respect to other features in this micro-area, it is not always possible to come to any conclusions concerning the origin of their development, though contact can be assumed for some cases. The calquing of the expression for 'because', *ɨamɨ goːmi*, in W.Ėven (specifically in the Lamunkhin dialect) from Sakha *toɣo dieteχχe* has already been mentioned above; in both languages, a literal translation of the construction is "if one were to say why". A case of plausible contact-induced development is the

presence of pronominal quotes in Sakha, W.Ėven, and E.Evenki (cf. Section 3.3.1, (15)); here, however, the direction of contact is unclear. The origin of the purposive construction in W.Ėven and E.Evenki (Section 3.3.3) is rather more complicated and cannot be elucidated with certainty. On the one hand, in both North Tungusic lects there is a purposive construction making use of an inherited purposive converb redundantly augmented by SAY (cf. (5) above); since the purposive converb is clearly a Tungusic feature which is lacking in Sakha, this might appear to be a North Tungusic innovation. However, the W.Ėven 3rd person purposive expresses 3rd person imperatives, thus making the W.Ėven construction identical to the Sakha non-coreferential purposive construction with 3rd person subordinate subject, which Sakha inherited from its Turkic ancestor. While the details are too complex to elaborate here, the most plausible explanation is that both E.Evenki and W.Ėven innovated purposive constructions marked by non-canonical SAY under Sakha influence, though in separate developments (cf. Pakendorf 2013).

Of special interest in this micro-area is Dolgan, the closest Turkic relative of Sakha, which is sometimes classified as a dialect of the latter (e.g. Voronkin 1999). Notwithstanding this close relationship, Dolgan has a considerably smaller functional load of non-canonical SAY than its sister language (cf. Table 4 and Appendix 2), and it even seems to have lost some of the inherited Turkic functions, such as Nam_Call. This might be due to the influence of languages lacking non-canonical SAY during the formation of Dolgan, in particular Western Evenki, with possible further influence from Samoyedic Nganasan (cf. Stachowski 1993: 15ff, 2010, Stapert, in preparation), and would thus represent a case of contact-induced feature loss.

6. Conclusions

We have shown that the text-based method of assessing areal and genealogical features applied in this paper has a number of advantages over the traditional grammar-based method, especially when differences between genres are minimized. The material basis for the comparison is much broader: many of the features we were able to take into account are not mentioned in standard descriptions, and syntactic and semantic properties of constructions are often observable only when extracted out of natural discourse. Furthermore, frequency data, which enable computational data analysis such as the Correspondence Analysis used here, can only be obtained through text

counts. In addition, the inclusion of data from closely related languages (Buryat and North Mongolian, Dolgan and Sakha) and dialects of Evenki and Ėven has enabled us to gain important insights into the distribution of non-canonical SAY in Siberia. This has demonstrated the considerable heterogeneity of the area that might otherwise have escaped notice.

However, even with these improved methodological tools, it is not always possible to unravel the complex factors underlying areal processes. While we have not been able to come up with definitive solutions for all problems, it has become clear that the spread of non-canonical SAY over the vast area of Siberia has been the result of multiple causes: demonstrated inheritance of certain functions in the Turkic and Mongolic languages coupled with diverse contact-induced developments and also independent innovations, most easily and clearly demonstrable for the origin of non-canonical SAY in Enets. The case of Dolgan shows that language contact need not always result in the gain of features, but can occasionally also lead to their loss, if these are absent in the contact language(s). In response to the question with which we began our study, namely whether the distribution of non-canonical SAY in Siberia is explicable in terms of common inheritance, language contact, or independent developments, we can thus state: it is all three. The interplay of these factors has resulted in the variegated picture we have identified: what at first sight appears to be a monolithic linguistic area turns out to be a set of interconnected micro-areas.

We have also been able to elucidate the borders of this conglomerate of micro-areas in the east and the west. In the east, the Siberian SAY-area reaches the Pacific coast in the south (Eastern Evenki, Udihe, Nanai, and Nivkh); an unexpected non-SAY island in this area is Negidal. Further north, it is flanked by non-SAY languages of the eastern Pacific (eastern Ėven dialects and Chukotko-Kamchatkan languages; cf. Section 3.1). In the west, the Siberian SAY-area extends into non-SAY languages on the Taimyr Peninsula in the far north, where it is represented by Dolgan and Enets. This extension is partly due to the recent migration of the Turkic-speaking Dolgans into the area, which would otherwise be more homogenous in its lack of non-canonical SAY. Enets constitutes an unexpected outlier, having developed non-canonical SAY through independent developments, as discussed in Section 5.1. In the southwest, the non-canonical SAY area of Siberia blends into the Turkic and Turkic-influenced region spreading from Central Asia to the Balkans (e.g. Pokrovskaja 1978: 156ff, Johanson 2002: 137, Erdal 2004: 488ff, Khanina 2007, Straughn 2008), following the

movement of Turkic-speaking peoples along the steppe belt of Eurasia. The limits of the spread of the SAY-area to the south are much less clear, as indicated in Section 2.1, where we assumed that at least some Siberian SAY-micro-areas may have connections with the well-known SAY-areas in East and South Asia. Unfortunately, however, other than descriptions of fairly basic functions, such as quote marker, complementizer, and marker of purpose and reason adjuncts (Ebert 1991, Masica 1993: 402-403, Chappell 2008, Genetti 2011: 58), we lack information on the scope of non-canonical SAY in these regions. A detailed text-based comparison of forms and functions, similar to that presented in this paper, would doubtlessly reveal a number of interesting areal connections spread over the greater part of the huge Eurasian landmass. This, however, remains a task for future research.

Abbreviations

ABL ablative
ACC accusative
ACT actor
ADJ adjective
ADV adverb
ALL allative
ALN alienable
ATTR attributive
BND bound
CAUS causative
CLOC cis-locative
CNJ conjoining
COND conditional
COP copula
CSE causee
CTR contrastive
CVB converb
DAT dative
DES designative
DIM dimunitive
DIR directive
DIV diversative
DS different subject
EMP emphatic
EV evidential
FOC focus
FREQ frequentative
FUT future
GEN genitive
GNR generic
HAB habitual
HORT hortative
HST hesitative
IDEO ideophone
IMP imperative
IMPS impersonal
IND indicative
INDIR indirect
INST instrumental
IPF imperfective
ITER iterative
LIM limitative
MDL modal
MED medium/middle
MUL multiplicative
NEG negation
NFUT non-future
NPST non-past
OBL oblique
PASS passive
PF perfective
PL plural
POSS possessive
PRED predicative
PROG progressive
PROL prolative
PROP proprietive
PRS present
PST past
PTC participle
PTL particle
PURP purposive
Q question
REC recent
REFL reflexive
SBEN subject beneficiary
SEQ sequential
SG singular
SIM simultaneous
SML similitive
SS same subject
TERM terminative
TR transitive
VR verbaliser
0 epenthetic vowel

?? unclear

References:

Abbott, Barbara. 2003. Some notes on quotation. *Belgian Journal of Linguistics* 17: 13-26.

Adelaar, William F.H. 1990. The role of quotations in Andean discourse. In: Harm Pinkster & Inge Genee (eds.): *Unity in diversity. Papers presented to Simon C. Dik on his 50th birthday*. Dordrecht: Foris.

Aikhenvald, Alexandra. 2004. *Evidentiality*. Oxford: OUP.

Aikhenvald, Alexandra. 2008. Semi-direct speech: Manambu and beyond. *Language Sciences* 30: 383-422.

Aikhenvald, Alexandra. 2009. Semantics and grammar in clause linking. In: Robert M.W. Dixon & Alexandra Aikhenvald (eds.), *The semantics of clause-linking. A retrospective*, 380-402. Oxford: Oxfor University Press.

Anderson, Gregory D. S. 2004. The languages of Central Siberia. Introduction and Overview. In: Edward Vajda (ed.), *Languages and prehistory of central Siberia*, 1-123. Amsterdam: Benjamins.

Anderson, Gregory D. S. 2006. Towards a Typology of the Siberian Linguistic Area. In: Yaron Matras, April McMahon & Nigel Vincent (eds.), *Linguistic areas. Convergence in historical and typological perspective*, 266-300. Basingstoke: Palgrave Macmillan.

Anderson, Gregory D. S. & K. David Harrison. 1999. *Tyvan*. München: Lincom Europa.

Avrorin, Valentin A. 1961. *Grammatika nanajskogo jazyka*, vol. 2. Moskva-Leningrad: Nauka.

Avrorin, Valentin A. 1986. *Materialy po nanajskomu jazyku i fol'kloru*. Leningrad: Nauka.

Avrorin, Valentin A. 2000. *Grammatika man'čurskogo pis'mennogo jazyka*. Sankt-Peterburg: Nauka.

Aydemir, Ahmet. 2009. *Converbien im Tuwinischen*. Wiesbaden: Harrassowitz.

Benbaji, Yitzhak. 2003. Who needs semantics of quotation marks? *Belgian Journal of Linguistics* 17: 27-49.

Bergelson, Mira B. & Andrej A. Kibrik. 1995. The system of switch-reference in Tuva: Converbal and masdar-case forms. In: Martin Haspelmath & Ekkehard König (eds.), *Converbs in cross-linguistic perspective*, 373-414. Berlin: Mouton de Gruyter.

Bertagaev, Trofim A. & Cybikžap B. Cydendambaev. 1962. *Grammatika burjatskogo jazyka. Sintaksis*. Moskva: IVL.

Bibiko, Hans-Jörg. 2005. WALS – The Interactive Reference Tool. Available on CD-ROM accompanying Haspelmath et al. 2005, and on http://wals.info

Binnick, Robert I. 1979. *Modern Mongolian. A transformational syntax*. Toronto: University of Toronto Press.

Bisang, Walter. 1992. *Das Verb im Chinesischen, Hmong, Vietnamesischen, Thai und Khmer*. Tübingen: Narr.

BNS 2000 = *Burjatskie narodnye skazki o životnyx. Bytovye*. Novosibirsk: Nauka.

Bogoras, Waldemar. 1917. *Koryak texts*. Leiden: Brill.

Boye, Kasper & Peter Harder. 2007. Complement-taking predicates. Usage and linguistic structure. *Studies in Language* 31: 569–606.

Broadwell, George A. 1991. Speaker and self in Choctaw. *International Journal of American Linguistics* 57: 411-425.

Brodskaja, Larisa M. 1987. Zavisimoe upotreblenie konstrukcii s prjamoj reč'ju v evenkijskom jazyke. In: Čeremisina (ed.), 51-72.

Brodskaja, Larisa M. 1988. *Složnopodčinenoe predloženie v evenkijskom jazyke*. Novosibirsk: Nauka.

Bulatova, Nadežda Ja. 1987. *Govory evenkov Amurskoj Oblasti*. Leningrad: Nauka.

Čeremisina, Majja I. 1987. Služebnoe slovo dep i formiruemye im sintaksičeskie konstrukcii (na materiale altajskogo jazyka). In: Čeremisina (ed.), 4-31.

Čeremisina, Majja I. (ed.) 1987. *Pokazateli svjazi v složnom predloženii na materiale jazykov raznykh system*. Novosibirsk: Nauka.

Čeremisov, Konstantin M. 1973. *Burjatsko-russkij slovar'*. Moskva: Sovjetskaja Enciklopedija.

Chappell, Hilary. 2008. Variation in the grammaticalization of complementizers from *verba dicendi* in Sinitic languages. *Linguistic Typology* 12: 45-98.

Chisarik, Erika & Wim van der Wurff. 2003. From 'say' to 'because': Grammaticalisation and reanalysis. Paper presented at the *Conference on Comparative Diachronic Syntax*, Leiden. http://www.let.leidenuniv.nl/ulcl/events/compdiachr/programme.htm

Clark, Larry. 1998. *Turkmen reference grammar*. Wiesbaden: Harrassowitz.

Collins, Chris & Phil Branigan. 1997. Quotative inversion. *Natural Language and Linguistic Theory* 15: 1–41.

Creissels, Denis. 2001. Setswana ideophones as uninflected predicative lexemes. In: F. K. Erhard Voeltz & Christa Kilian-Hatz (eds.), *Ideophones*, 75–85. Amsterdam: Benjamins.

Dehé, Nicole & Anne Wichmann. 2010. Sentence-initial *I think (that)* and *I believe (that)*. Prosodic evidence for use as main clause, comment clause and discourse marker. *Studies in Language* 34: 36-74.

Ebert, Karen H. 1986. Reported speech in some languages of Nepal. In: Florian Coulmas (ed.), *Direct and indirect speech,* 145–160. Berlin: Mouton de Gruyter.

Ebert, Karen H. 1991. Vom Verbum dicendi zur Konjunktion – Ein Kapitel universaler Grammatikalisierung. In: Walter Bisang & Peter Rinderknecht (eds.), *Von Europa bis Ozeanien - von der Antonymie zum Relativsatz, Gedenkschrift für Meinrad Scheller*. Arbeiten des Seminars für Allgemeine Sprachwissenschaft 11, 77-95. Zürich: Universität Zürich.

Erdal, Marcel (2004) *A Grammar of Old Turkic*. Leiden: Brill.

Fil'chenko, Andrey. 2007. *Documentation of Eastern Khanty*. http://www.policy.hu/filtchenko/Documenting Eastern Khanty/index.htm

FD 2000 = *Fol'klor dolgan.* Novosibirsk: Nauka.

Fried, Robert Wayne. 2010. *A grammar of Bao'an Tu*. Dissertation, University of Buffalo. http://gradworks.umi.com/3407976.pdf

Genetti, Carol. 2011. Direct speech reports and the cline of prosodic integration in Dolakha Newar. *Himalayan Linguistics* 10: 54-76.

Göksel, Aslı & Celia Kerslake. 2005. *Turkish: A comprehensive grammar.* Oxford: Routledge.

Gorelova, Liliya M. 2002. *Manchu grammar*. Leiden: Brill.

Güldemann, Tom. 2002. When 'say' is not say: the functional versatility of the Bantu quotative marker ti with special reference to Shona. In: Tom Güldemann & Manfred von Roncador (eds.), *Reported discourse: A meeting ground for different linguistic domains*, 253-287. Amsterdam: Benjamins.

Güldemann, Tom. 2008. *Quotative indexes in African languages: A synchronic and diachronic survey*. Berlin: Mouton de Gruyter.

Harrison, K. David. 2005. A Tuvan hero tale. *Journal of the American Oriental Society* 125: 1-30.

Heine, Bernd & Tania Kuteva. 2002. *World lexicon of grammaticalization*. Cambridge: Cambridge University Press.

Heine, Bernd & Tania Kuteva. 2005. *Language contact and grammatical change*. Cambridge: Cambridge University Press.

Helimski, Eugene. 2003. Areal groupings within and across the borders of the Uralic language family. *Nyelvtudományi közlemények* 100: 156-167.

Hooper, Joan & Sandra Thompson. 1973. On the applicability of root transformations. *Linguistic Inquiry* 4: 465–497.

JaNS 2008 = *Jakutskie narodnye skazki*. Novosibirsk: Nauka.

Janhunen, Juha. 2003. Written Mongol. In: Janhunen (ed.), 30-56.

Janhunen, Juha (ed.). 2003. *The Mongolic languages*. London: Routledge.

Johanson, Lars. 1998a. The structure of Turkic. In: Lars Johanson & Eva Csató (eds.), *The Turkic languages*, 30-66. London: Routledge.

Johanson, Lars. 1998b. The history of Turkic. In: Lars Johanson & Eva Csató (eds.), *The Turkic languages*, 81-125. London: Routledge.

Johanson, Lars. 2002. *Structural factors in Turkic language contacts*. Richmond: Curzon.

Jochel'son, Vladimir I. 1900. *Materialy po izučeniju jukagirskogo jazyka i fol'klora*. Sankt-Peterburg: Imperatorskaja Akademija Nauk.

Kálmán, Béla. 1976. *Wogulische Texte mit einem Glossar*. Budapest: Akadémiai Kiadó.

Kazakevič, Ol'ga A., Elena M. Budjanskaja, Julija E. Galjamina. 2008. Istorii žizni avtokhtonnogo naselenija Sibiri: publikacija glossirovannykh tekstov. Publikacija pervaja. *Vestnik RGGU 6/08. Serija Jazykoznanie/ Moskovskij lingvističeskij žurnal* 10: 246-285. Moskva: Izdatel'stvo RGGU.

Khabtagaeva, Baiarma. 2006. The intensitity of the Mongolian influence on the Tuva language. *Acta Orientalia* 59: 85-94.

Khanina, Olesya. 2007. Konstrukcii s grammatikalizovannym konverbom glagola reči. In: Ekaterina A. Ljutikova, Konstantin I. Kazenin, Sergej G. Tatevosov & V. D. Solovjev (eds.), *Mišarskij dialekt tatarskogo jazyka. Očerki po sintaksisu i semantike,* 126-140. Kazan: Magarif.

Khanina, Olesya (MS), Sentencial'nye aktanty čuvašskogo jazyka.

Khasanova Marina & Aleksandr Pevnov. 2003. *Mify i skazki negidal'cev*. Osaka: ELPR Publications.

Kibrik Aleksandr E., Sandro V. Kodzasov & Irina A. Muravyova. 2004. *Language and folklore of the Alutor people*. Osaka: ELPR Publications.

Klamer, Marian. 2000. How report verbs become quote markers and complementizers. *Lingua* 110: 69-98.

Kornfilt, Jaklin. 1997. *Turkish.* London: Routledge.

Kornfilt, Jaklin. 2001. On the syntax and morphology of clausal complements and adjuncts in the Turkic languages. In: Walter Bisang (ed.), *Aspects of typology and universals,* 63-82. Berlin: Akademie Verlag.

Krejnovič, Erukhim A. 1979. Nivkhskij jazyk. In: *Jazyki Azii i Afriki* 3, 295-329. Moskva: Nauka.

Krjukova, Elena A. & Pavel Ju. Glazunov. 2010. Ketskie teksty. In: Andrey Ju. Fil'čenko (ed.), *Annotirovannye fol'klornye teksty obsko-jenisejskogo jazykovogo areala*, 184-210. Tomsk: Veter.

Kurilov, Gavril N. 2005. *Fol'klor jukagirov.* Novosibirsk: Nauka.

Kuz'mina, Raisa P. 2010. *Jazyk lamunkhinskikh évenov*. Novosibirsk: Nauka.

Lessing, Ferdinand D. 1960. *Mongolian-English dictionary*. Berkeley: UC Press.

Malchukov, Andrej. 2006. Yakut interference in North-Tungusic languages. In: Hendrik Boeschoten & Lars Johanson (eds.) *Turkic languages in contact*, 122-139. Wiesbaden: Harrassowitz.

Malchukov, Andrej. 2008. *Sintaksis evenskogo jazyka*. Sankt-Peterburg: Nauka.

Martin, Samuel E. 1961. *Dagur Mongolian grammar, texts, and lexicon*. The Hague: Mouton.

Masica, Colin. 1993. *The Indo-Aryan languages*. Cambridge: CUP.

Maslova, Elena. 2001. *Yukaghir texts*. Wiesbaden: Harrassowitz.

Matisoff, James. 1991. Areal and universal dimensions of grammatization in Lahu. In: Elizabeth C. Traugott & Bernd Heine (eds.), *Approaches to grammaticalization*, vol. 2, 383-454. Amsterdam: Benjamins.

Mattissen, Johanna. 2008. Converbs in Nivkh. In: Karen H. Ebert, Johanna Mattissen & Rafael Suter (eds.), *From Siberia to Ethiopia. Converbs in a cross-linguistic perspective*. Arbeiten des Seminars für Allgemeine Sprachwissenschaft 20, 91-124. Zürich: Universität Zürich.

McGregor, William. 1994. The grammar of reported speech and thought in Gooniyandi. *Australian Journal of Linguistics* 14: 63-92.

MDKL = *Multimedia Database of Ket Language*. http://minlang.srcc.msu.ru/kets/g/g_qtext_en.htm

Mongush, Dorgu-ool Aldyn-oolovič. 1987. O nekotoryx modeljax predloženij v tuvinskom jazyke, obrazovannyx pri učastii proizvodnyx ot glagola de=/di= "govorit'", "skazat'". In: Čeremisina (ed.), 83-88.

Muniev, Bembe D. 1977. *Kalmycko-russkij slovar'*. Moskva: Russkij jazyk.

Nedeljaev, V. M.; D. M. Nasilov D.M.; A. M. Scherbak & E. R. Tenishev. 1969. *Drevnetjurkskij slovar'*. Leningrad: Nauka.

Nikolaeva, Irina. 1999. *Ostyak texts in the Obdorsk dialect*. Wiesbaden: Harrassowitz.

Nikolaeva, Irina. 2004. *Kolyma Yukaghir documentation*. http://www.sgr.fi/yukaghir/index.html

Nikolaeva, Irina. 2005. Review of *A grammar of Kolyma Yukaghir*, by Elena Maslova. *Linguistic Typology* 9: 299–325.

Nikolaeva, Irina. 2006. *A historical dictionary of Yukaghir*. Berlin: de Gruyter.

Nikolaeva, Irina; Elena Perekhvalskaya & Maria Tolskaya. 2003. *Udihe texts*. Osaka: ELPR.

Nikolaeva, Irina & Maria Tolskaya. 2001. *A grammar of Udihe*. Berlin: Mouton de Gruyter.

NMV 1974 = *Nordmongolische Volksdichtung*. Gesammelt von G.J. Ramstedt, bearbeitet, übersetzt und herausgegeben von Harry Halén. Band II. Helsinki: Suomalais-Ugrilainen Seura.

Noonan, Michael. 2006. Direct speech as a rhetorical style in Chantyal. *Himalayan Linguistics* 6: 1-32.

Pal'mbax, Aleksandr A. 1955. *Tuvinsko-russkij slovar'*. Moskva: GIINS.

Pakendorf, Brigitte. 2009. Intensive contact and the copying of paradigms – an Ėven dialect in contact with Sakha (Yakut). *Journal of Language Contact* 2: 85-110.

Pakendorf, Brigitte. 2013. Incipient grammaticalization of a redundant purpose clause marker in Lamunxin Ėven: contact-induced change or independent innovation? In: Robbeets, Martine & Hubert Cuyckens (eds.), *Shared grammaticalization. With special focus on the Transeurasian languages*, 259-283. Amsterdam: Benjamins.

Panfilov, Vladimir Z. 1965. *Grammatika nivkhskogo jazyka*, vol. 2. Moskva-Leningrad: Nauka.

Pokrovskaja Ljudmila A. 1978. *Sintaksis gagauzskogo jazyka v sravnitel'nom osveščenii*. Moskva: Nauka.

Poppe, Nikolaj N. 1937. *Grammatika pis'menno-mongolskogo jazyka*. Moskva: IAN SSSR.

Reesink, Ger P. 1993. Inner speech in Papuan languages. *Language and Linguistics in Melanesia* 24: 217-25.

Romanova, Agnija V. & Anna N. Myreeva. 1962-1964-1968. *Dialektologičeskie materialy po govoram evenkov Jakutskoj ASSR, 1-3*. Moskva–Leningrad: AN SSSR.

Sanžeev, Garma D. 1964. *Sravnitel'naja grammatika mongol'skikh jazykov*. Moskva: Nauka.

Saxena, Anju. 1988. On syntactic convergence: The case of the verb 'say' in Tibeto-Burman. In: *Berkeley Linguistic Society (BSL)* 23, 375–388.

Saxena, Anju. 1995. Unidirectional grammaticalization: Diachronic and synchronic evidence. *Sprachtypologie und Universalienforschung* 48: 350-372.

Schönig, Claus. 2003. Turko–Mongolic relations. In: Janhunen (ed.), 403-419.

Sechenbaatar, Borjigin. 2003. *The Chakhar dialect of Mongol.* Helsinki: Finno-Ugrian Society.

Serdobol'skaja, Natal'ja V. & Svetlana Ju. Toldova. 2006. Puti grammatikalizacii glagola reči *manaš* v konstrukcijakh s sentencijal'nymi aktantami v marijskom jazyke. Paper presented at LENCA 3. http://www.ling.helsinki.fi/uhlcs/LENCA/LENCA-3/information/abstracts.html

Shiraishi, Hidetoshi & Galina Dem'ianovna Lok. 2008-2009. *Sound Materials of the Nivkh Language, 5-6*. http://ext-web.edu.sgu.ac.jp/hidetos/

Skribnik, Elena. 1987. Burjatskie polipredikativnye konstrukcii s pokazatelem *geže*. In: Čeremisina (ed.), 32-50.

Skribnik, Elena. 2003. Buryat. In: Janhunen (ed.), 102-28.

Slater, Keith W. 2003. *A grammar of Mangghuer*. London: Routledge Curzon.

Slepcov, P.A. (ed.). 1972. *Jakutsko-russkij slovar'*. Moskva: Izdatel'stvo Sovetskaja Enciklopedija.

Stachowski, Marek. 1993. *Dolganischer Wortschatz*. Krakow: Nakł. Uniewersytetu Jagiellońskieg

Stachowski, Marek. 2010. On the article-like use of the Px2Sg in Dolgan, Nganasan and some other languages in an areal Siberian context. *Finnisch-Ugrische Mitteilungen* 32/33: 587-93.

Stapert, Eugénie (in preparation) *Contact-induced change in Dolgan: An interdisciplinary investigation into the role of linguistic data for the reconstruction of a people's (pre)history*. PhD dissertation to be submitted to Leiden University.

Straughn, Christopher A. 2008. History vs. theory: The grammaticalization of the Uzbek say-complementizers. *Poznań Linguistic Meeting* 39 [http://home.uchicago.edu/~straughn]

Suñer, Margarita. 2000. The Syntax of direct quotes with special reference to Spanish and English. *Natural Language and Linguistic Theory* 18: 525-578

Svantesson, Jan-Olof. 2003. Khalkha. In: Janhunen (ed.), 154-176.

TNS 1994 = *Tuvinskie narodnye skazki*. Novosibirsk: Nauka.

Tugolukov, Vladilen A. 1997. Etnogenez i etničeskaja istorija. In: V. A. Turaev et al. (eds.) *Istorija i kul'tura evenov*, 13-57. Sankt-Peterburg: Nauka.

Varlamova, Galina I. & Aleksandr N. Varlamov. 2004. *Skazanija vostočnyx evenkov*. Yakutsk: SO RAN.

Vasilevič Glafira M. 1936. *Sbornik materialov po evenkijskomu (tungusskomu) fol'kloru*. Leningrad: Izdatel'stvo instituta narodov several CIK SSSR im. P.G. Smidoviča.

Vasilevič, Glafira M. 1948. *Očerki dialektov evenkijskogo jazyka*. Leningrad: Nauka.

Vasilevič Glafira M. 1966. *Istoričeskij fol'klor evenkov*. Moskva–Leningrad: Nauka.

Voronkin, Mixail S. 1999. *Dialektnaja sistema jazyka saxa.* Novosibirsk: Nauka.

Weiers, Michael. 2003. Moghol. In: Janhunen (ed.), 248-264.

Werner, Heinrich. 1997. *Die ketische Sprache*. Wiesbaden: Harrassowitz.

Wurm, Stephen A. 1996. Indigenous lingue franche and bilingualism in Siberia (beginning of the twentieth century). In: Stephen Wurm, Peter Mühlhäusler, & Darrell Tryon (eds.), *Atlas of languages of intercultural communication in the Pacific, Asia, and the Americas*, 975-8. Berlin: Mouton de Gruyter.

Yakup, Abdurishid. 2005. *The Turfan dialect of Uyghur*. Wiesbaden: Harrassowitz.

Žukova, Alevtina N. 1980. *Jazyk palanskikh korjakov*. Leningrad: Nauka.

Appendix 1: Data sources

Language family	Language/dialect		Source	Size
Yeniseic	Ket		Kazakevič et al. 2008; MDKL; Krjukova & Glazunov 2010	~3,500 words
Uralic	Nganasan		field data Maria Brykina & Valentin Gusev (http://www.iling-ran.ru/gusev/Nganasan)	~5,750 words
	Enets		field data Andrej Shluinsky	~4,100 words
	Khanty		Nikolaeva 1999; Fil'chenko 2007	~13,100 words
	Mansi		Kálmán 1976	~8,000 words
Mongolic	North Mongolian		NMV 1974	~8,000 words
	Buryat		BNS 2000	~10,000 words
Turkic	Sakha	LHist	field data BP	~30,600 words
		Folk	JaNS 2008	~5,000 words
	Dolgan	LHist	field data Eugénie Stapert	~14,900 words
		Folk	FD 2000	~5,000 words
	Tuvan		TNS 1994; Harrison 2005	~8,500 words
	Shor		field data Irina Nevskaya, Shorica (http://shoriya.ngpi.rdtc.ru/)	~10,500 words
North Tungusic	Western Ėven	LHist	field data BP & DM	~48,300 words
		Folk	field data BP & DM; Kuz'mina 2010	~9,300 words
	Eastern Ėven		field data DM, BP, & Alexandra Lavrillier	~51,700 words
	Western Evenki		Vasilevič 1936, Vasilevič 1966	~16,100 words
	Eastern Evenki		Varlamova & Varlamov 2004, Vasilevič 1948, Romanova & Myreeva 1964, Bulatova 1987	~11,800 words
	Negidal		Khasanova & Pevnov 2003	~8,000 words
South Tungusic	Udihe		Nikolaeva, Perekhvalskaya & Tolskaya 2003	~6,800 words
	Nanai		Avrorin 1986, field data Sofia Oskolsakaja	~7,500 words
Yukaghir	Kolyma Yukaghir		Maslova 2001, Nikolaeva 2004	~20,600 words
Chukotko-Kamchatkan	Koryak		Bogoras 1917, Žukova 1980	~4,400 words
	Alutor		Kibrik, Kodzasov & Muravyova 2004	~8,200 words
Isolate	Nivkh		Shiraishi & Lok 2008-2009	~8,100 words

Appendix 2: Numerical data on non-canonical SAY

	Enets	NMong	Buryat	Tuvan	Shor	Dolgan LifeHist	Dolgan Folklore	Sakha LifeHist	Sakha Folklore	WĖven LifeHist	WĖven Folklore	EEvenki	KYukagh	Nanai	Udihe	Nivkh
Qu_DiffV		128 (35.3)	151 (25.2)	84 (21.6)	32 (17.8)	4 (1.1)	19 (9.5)	28 (4.2)	27 (9.9)	41 (4.4)	16 (6.1)	15 (2.8)	9 (1.1)	1 (0.4)	5 (1.8)	2 (2.7)
Qu_SameV			2 (0.4)	4 (1.0)	1 (0.6)	5 (1.3)	1 (0.5)	2 (0.3)		14 (1.5)	4 (1.5)	6 (1.1)	16 (2.0)	21 (8.4)		
Qu_Noun		1 (0.3)	3 (0.5)	15 (3.9)	1 (0.6)		4 (2.0)	5 (0.7)	5 (1.8)	1 (0.1)	1 (0.4)		1 (0.1)			
Qu_NoV		2 (0.5)	27 (4.5)	6 (1.5)	[+][1]	3 (0.8)		16 (2.4)	7 (2.6)	59 (6.3)	8 (3.0)	1 (0.2)	7 (0.9)	4 (1.6)	15 (5.4)	
Pron_Qu					4 (2.2)			3 (0.5)	2 (0.7)	5 (0.5)	5 (1.9)	3 (0.6)				
Comp_SPCH		2 (0.6)		2 (0.5)	[+]		1 (0.5)			7 (0.8)	1 (0.4)		6 (0.7)			
Comp_COGN		4 (1.1)	4 (0.7)	3 (0.8)	4 (2.2)	2 (0.5)	1 (0.5)	17 (2.6)		25 (2.7)	3 (1.1)	[+][2]	10 (1.2)	2 (0.8)	2 (0.7)	
Comp_EMOT			1 (0.2)		[+]	1 (0.3)				10 (1.1)	1 (0.4)	2 (0.4)	14 (1.7)			
Comp_PERC		1 (0.3)		[+][3]	6 (3.3)				1 (0.4)	2 (0.2)			7 (0.9)		[+][4]	
Adj_PURP		16 (4.4)	5 (0.8)	11 (2.8)	2 (1.1)		1 (0.5)	21 (3.2)	3 (1.1)	55 (5.9)	3 (1.1)	4 (0.8)	5 (0.6)			
Adj_REAS		6 (1.6)	2 (0.3)	1 (0.3)	[+]	3 (0.8)	1 (0.5)	14 (2.1)	3 (1.1)	12 (1.3)	7 (2.7)	2 (0.4)	6 (0.7)			
Adj_CONC								6 (0.9)								
Adj_COND	6 (4.9)															
Evid_HSY	18 (14.7)	1 (0.3)	[+][5]				1 (0.5)	2 (0.3)	1 (0.4)	3 (0.3)				2 (0.8)	4 (1.5)	4 (5.4)

[1] All additional attestations for Shor: Irina Nevskaya, p.c.

[2] Attestations in Brodskaja (1988: 72-3).

[3] Attestations in Anderson & Harrison (1999: 75).

[4] Attestations in Nikolaeva & Tolskaja (2001: 662).

[5] Attestations in Skribnik (2003: 119).

Evid_Perc/Mir														4 (1.6)	18 (6.5)	
Nam_CallV			[+][6]		3 (1.7)			12 (1.8)	1 (0.4)	9 (1.0)	3 (1.1)	1 (0.2)	8 (1.0)		3 (1.1)	
Nam_DiffV								6 (0.9)		1 (0.1)						
Nam_Pred		2 (0.6)					2 (1.0)	34 (5.2)	2 (0.7)	5(0.5)		1 (0.2)				
Nam_App		11 (3.0)	3 (0.5)	14 (3.6)	5 (2.8)	7 (1.9)		38 (5.8)	13 (4.8)	2 (0.2)	1 (0.4)	11 (2.1)	7 (0.9)			
Nam_Term				1 (0.3)		2 (0.5)		96 (14.6)		14 (1.5)	1 (0.4)	10 (1.9)	1 (0.1)			
Nam_Cl/Rel		1 (0.3)	3 (0.5)	5 (1.3)				17 (2.6)		1 (0.1)			2 (0.2)			
Disc_Ptl		10 (2.8)	64 (10.7)	5 (1.3)		23 (6.2)		52 (7.9)	6 (2.2)	7 (0.8)		8 (1.5)	1 (0.1)			
Disc_ENUM			[+][7]	1 (0.3)				31 (4.7)	3 (1.1)							
Disc_Top		1 (0.3)	1 (0.2)	13 (3.3)		1 (0.3)		14 (2.1)		6 (0.6)		[+][8]				
Compr										2 (0.2)						
Lex_Caus		[+][9]	5 (0.8)	2 (0.5)				6 (0.9)		13 (1.4)						
Lex_Other		1 (0.3)	2 (0.3)	1 (0.3)	13 (7.2)											
Lex_Par						1 (0.3)				12 (1.3)						
Aux		14 (3.9)	13 (2.2)	8 (2.1)												
Light_V		(?)[10]	9 (1.5)	3 (0.8)												

[6] Attestations in Skribnik (1987: 32).

[7] Attestations in Bertagaev & Cydendambaev (1962: 223ff.)

[8] Attestations in Vasilevič (1948: 271).

[9] Attested in Lessing (1960: 451).

[10] According to Sechenbaatar (2003:153), light verb function is frequent in one of the central dialects of Mongolian (Chakhar). It is plausible to assume that it is also present in the northern dialects, from which we draw our data.

Unclear	1 (0.7)	4 (1.1)	9 (1.5)	8 (2.1)			1 (0.5)	18 (2.7)	1 (0.4)	15 (1.6)	2 (0.8)	2 (0.4)	8 (1.0)			

absolute number of tokens (% of all instances of SAY); [+]: not attested in texts, only in secondary literatur