Wagner-Nagy Beáta, Szeverényi Sándor

LINGUISTICALLY ANNOTATED SPOKEN NGANASAN CORPUS

The paper discusses the key issues of the annotation method employed in the project "Linguistically annotated spoken Nganasan corpus". The data are processed and stored in the EXMARaLDA format. The annotation of the database involves grammatical and part-of-speech tagging (made in Toolbox or Flex), translation into Russian and English. However, the present paper addresses the questions of syntactic roles, and information structure. For this purpose we use the format designed by other researchers and adapted by us to the Nganasan language. In the paper we describe the system of annotation (tags, terms and their clarification) illustrated by a large amount of Nganasan examples.

Key words: *Nganasan, annotation, corpus, endangered language, language documentation.*

1. The Spoken Nganasan Corpus

The Spoken Nganasan Corpus has been created as part of a project[1] whose goal is to generate a digital, machine-searchable corpus of spoken Nganasan and, based on this corpus, to prepare a corpus-based reference grammar of the language. This project will fill basic gaps in the existing research into Nganasan descriptive grammar creating new and more widely accessible materials and information on this lesser known and severely endangered Uralic language. As Nganasan is not a completely unknown language and there is a considerable amount of language data and descriptions available, the focus of work is corpus building. The bulk of the language material to be integrated, glossed and annotated has been collected by several researchers[2] and is available in audio format, most of it also in video format. In the final version of the corpus the texts will be aligned with the audio/video files.

The transcription data as well as the metadata of the corpus are processed and stored in EXMARaLDA[3] format, which is both a well-documented and widely used XML-format. The data organization, archiving and publication process carried out by the Hamburg Centre for Language Corpora (HZSK) will convert the transcription files into other widely used formats for transcription (ELAN, Praat, etc.) as well as visualization and publication (Word/RTF, HTML, SVG, PDF), thus strengthening the reusability as well as sustainability of the corpus. The glossed and annotated communications (texts) are merged with the help of the EXMARaLDA Corpus Manager (Coma). The corpus created with Coma can be analyzed with the EXMARaLDA Analysis and Concordance-Tool (EXAKT)[4] (cf. Schmidt and Wörner, 2005; Wörner, 2010). With the help of EXAKT the annotated texts of the corpus can be searched in complex ways and the results listed, which makes it possible to discover new grammatical and other patterns. The program also allows the user to group and filter results as well as to compare them with metadata.

The metadata include information on the informants as well as the recorded communicative events. Metadata related to the informants include in all cases biographical information and the linguistic biography of the speaker. Further relevant data will also be included whenever it is available. Metadata on the communicative event will include interaction type, location and time, and language used.

Within the project period, sub-corpora will also be annotated for categories of syntactic functions, information structure, and thematic roles. The minimal requirements such a database should fulfill are English (and Russian) translations as well as English (and Russian) interlinear glosses.

The material prepared will be made available in the form of a searchable online text corpus accessible to the research community with a password protected online corpus facilitating multiple search

[1] Corpus based grammatical studies on Nganasan, supported by the DFG (German Research Grant).

[2] Maria Brykina, Valentin Gusev, Eugen Helimski, Jean-Luc Lambert, Tibor Mikola, Sándor Szeverényi, Beáta Wagner-Nagy in collaboration with other colleagues. We would like to thank here all the people who made this research possible.

[3] URL: http://www.exmaralda.org

[4] URL: http://www.exmaralda.org/tool/exakt/

options and concordance based analysis. Thus, the corpus will allow further research on the Nganasan language.

2. Annotation in EXMARaLDA

It is necessary that the data are morphologically glossed and tagged for parts of speech with Toolbox or Flex and also further annotated and processed with EXMARaLDA. For this, it has been necessary to create a software tool for converting the data from Toolbox and Flex to EXMARaLDA, which keeps tokenization in accordance with EXMARaLDA format. This work was done by Alexandre Archipov. At present the corpus contains 59 texts converted into the EXMARaLDA format and aligned with the corresponding audio files.

Annotation for every communicative event contains one tier of the type Transcription (T) for each speaker – this is the tier that contains the Nganasan text and aligned with the audio/video-files. In the tier reference (ref), the name of the communication and the number of the sentence is noted. The numbering is created automatically by Toolbox or Flex and imported into EXMARaLDA. This tier and the tier source text (st), if there is one, have a type description (d), which can be 'transcription', 'description', 'annotation' or 'comment'. All other tiers containing additional analytic information about the transcription have a type annotation (a). The tier tx is the line for interlinearization, which provides the basis for glossing in Flex or Toolbox. For the description of other tiers in EXMARaLDA, see Table 1 below.

Table 1

Tiers in EXMARaLDA in the Nganasan Spoken Corpus

TIERS	Comments	Type
ref	Name of the communication	d
st	Source texts: normally in Cyrillic transliteration	d
ts	Transcription (what is heard)	t
tx	Tier for interlinealization	a
mb	Morpheme break	a
mp	Morphophonemes, underlying forms	a
gr	Morphological annotation: Russian gloss for each morpheme	a
ge	Morphological annotation: English gloss for each morpheme	a
ps	Part of speech categorization for each morpheme	a
SeR	Annotation of semantic roles and syntactic functions	a
IST	Annotation of information status	a
fr	Free Russian transcription	a
fe	Free English transcription	a
nt	Notes on the text unit	

EXMARaLDA offers the possibility to insert a practically infinite number of tiers for annotation, which makes multiple-tier annotation easily doable. Of the numerous annotation possibilities, in the present paper we have chosen two tiers, which are described in Sections 3 and 4.

3. The annotation of thematic roles and syntactic functions

As for the annotation of syntactic functions and thematic roles, we have selected as sample the schema of GRAID: Grammatical Relations and Animacy in Discourse (cf. Haig and Schnell, 2011, 2014). GRAID 6 was developed for the annotation of endangered Oceanic languages, however, the new version, GRAID 7, has been applied to other languages (such as Semitic, Iranian, and English) as well, so it incorporates the experience collected in work with these languages as well. Given that all languages are different, the GRAID system cannot be readily applied to any new language – but this is

not necessary either, since every annotator can modify the GRAID scheme according to their needs and possibilities. Our goal with the GRAID-based analysis is, in addition to entering morphological information, to mark syntactic and semantic information in our corpus so that these types of information could be analyzed together in the data of the corpus. The annotation scheme applied to Nganasan has been developed on the basis of the GRAID system but it also differs from it. The Nganasan version of the GRAID system uses approximately 27 symbols. In EXMARaLDA the annotations not need to be carried out manually. In order to provide a unified system of annotation, we have created an annotation panel for thematic roles, which can be opened in EXMARaLDA's Partitur-Editor. This makes annotation faster and ensures its uniformity.

During the annotation, we take into account three factors: we annotate thematic roles and syntactic functions, and we provide information on their referents.

3.1. Annotation of thematic roles

3.1.1. Form of referent. In the corpus, the form of the referent is annotated. Not all possible factors of such forms are provided, but noun phrase and pronominal referents are differentiated. In annotating the thematic role Locative, whether the referent is adverbial, postpositional or nominal may also play a role. Similarly to GRAID, we also use the category <other>. It is used when we cannot or do not want to specify something at the present stage of annotation: for instance, some determiners are marked with the label <other> at present, but these will be easy to find and make more specific at a later stage. The categories which are used in specifying the form of the referent are listed in Table 2 below. Given that Nganasan is a pro-drop language, it is useful to mark whether the referent in question is expressed overtly in the sentence or not. Accordingly, for the form of referential expressions the following glosses are used:

Table 2

Form of referent

pro	np	n	0	adv	pp	other
Free pronoun	Noun phrase	Nominal	Deleted	Adverb	Postposition	Form which is not considered relevant

3.1.2. Properties of the referent. The inherent properties of the referent include the person. We annotate all three persons. Semantically the referent can be a human or non-human. Human referents are annotated with the symbol <h>, while the non-human referents are bare. In Nganasan the feature [±human] probably does not play a special role as far as thematic relations are concerned, however, we decided to include it in the annotation list anyway.

The properties of the referent are linked to the form categories with the symbol <.>. In contrast to GRAID, we do not annotate anthropomorphised discourse participants. However, if it turns out that there is a need for such a category, the annotation feature can be easily added.

3.2. Semantic roles

To this day there exists no unified list of semantic roles despite the fact that argument structure and the assignment of thematic roles are hot topics in the fields of semantics and syntax these days (cf. Dowty, 1989, 1991; Grimshaw, 1990; Butt, 2005; etc.). In our system of annotations, we have taken into account the thematic roles used in GRAID, but additionally we annotate some other semantic roles too, such as the recipient (R) and benefactor (B). It also has to be noted that certain thematic roles have not been differentiated yet: for instance, no differentiation is currently made between Agent and Experiencer. While in the sentence *Mary loves Peter* the argument *Mary* is an Experiencer as far as its thematic role is concerned due to the fact that it does not control the action, in our system the arguments of such sentences are categorized as Agents at the moment.

In the same way, we do not differentiate between a Patient and a Theme. While, for instance, in a sentence with a ditransitive verb the entity that is handed over by the Agent to the Recipient is the

Theme, in our system it is treated as if it were a Patient. The category of Recipient is annotated, however. Differentiating between a Recipient and a Goal is not unproblematic. One of the criteria for doing so is that if the verb expresses an actual or mental transfer, the argument at the other end is a Recipient. Naturally, the argument of verbs expressing a mental transfer is not a real recipient but only a recipient-like argument; this is not separately annotated (cf. Malchukov et al., 2010). Several other thematic roles have not been included in the list at present, such as Source and Undergoer but can be included at a later stage. For annotating the thematic roles, the following glosses are used[5]:

Table 3

Thematic roles

A(gent)	The initiator of the action
	The entity that experiences the action (actually, the experiencer)
P(atient)	The undergoer of the action
	The entity which is moved by some action (theme)
G(oal)	The location or entity in the direction of which something moves
L(ocation)	The locative argument of a verb, a place in which something is situated
R(ecipient)	The animate recipient of transfer, and addressee of verb of speech
B(enefactive)	The entity for whose benefit the action was performed
Poss(essor)	The possessor
Ins(trument)	The medium by which the action or event is performed

3.3. Syntactic functions

By annotating grammatical relations we focus only on the major syntactic functions as S, A and P, as well as on the predicate, which can be nominal or verbal, making this distinction necessary to differentiate as well. The verbal predicate is annotated as <v:pred>. The first element of the abbreviation refers to the type of predicate (nominal or verbal), whereas the second one to the role played in the sentence – that is, in this case, that the given verb functions as a predicate. There are, however, verbal predicates that go together with a copula which carries certain grammatical functions such as modal or tense marker, as sentence (1) shows. Here the actual predicate is annotated as a predicate, while the element bearing the tense marker receives the label copula.

(1) KES_061020_MyLife_nar.011

mb	*Bəhi͡a*	*i-sʲüə,…*
ge	bad.[3SG]	be-PST.[3SG]
#	n:pred	cop
fe	‘It was bad…’	

As the sentence above well demonstrates, a nominal element can play the role of a predicate. But Nganasan has the characteristic that even adjectives and particles can occur in this position (see Table 4 below).

Table 4

The form and function of the predicate

Form	v:pred	n:pred	adj:pred	ptcl:pred	cop	aux	aux.neg
Description	Verbal predicate	Nominal predicate	Attributive predicate	Particle predicate	Copula	Auxiliary	Negative auxiliary

In addition to purely verbal predicates, auxiliaries are also differentiated. In sentences that contain a structure with an auxiliary, the latter receives the annotation aux or aux.neg, whereas the connegative

[5] We rely on Gawron (2007) for defining thematic roles.

form of the main verb receives the label <v:pred>. The annotation scheme referring to the predicates is summarized in Table 4.

There are cases where one element has to be assigned two thematic or syntactic roles during annotation. This can happen, for instance, when the same word is marked for the recipient and the patient, or when a pro-drop phenomenon occurs during which the pronominal agent/subject is not expressed overtly. In the former case two thematic roles have to be marked, while in the latter two syntactic functions have to be annotated. In such cases the given cell is annotated for both functions or roles. The sentences in examples (2) and (3) below illustrate this. Example (2) provides a sentence in which the same word is marked for the Benefactor and the Patient, while (3) demonstrates a case where the subject is referred to only by the inflection on the verb. The latter is a frequent occurrence in Nganasan.

(2) ChND_061101_TwoTents_flkd.015

mb	*Maa-güə-ðə-mtə*	*ŋəði-ʔə-ŋ?*
ge	what-EMPH-DST-ACC.2SG	find-PF-2SG
#	pro:P/0.2.h:B	v:pred
fe	‘Did you find something for yourself?’	

We can see that the first word of the sentence is marked for two thematic roles: on the one hand, it is a pronominal Patient, but on the other hand it is also coded for the Benefactor. This is a frequent occurrence in Nganasan when the (pre)destinative suffix is used.

(3) KES_061020_MyLife_nar.002

mb	*d’esɨ-galʲi*	*bəðu͡a-suə-m*	*təbtə*
ge	father-PRIV.SG	grow(tr)-PST-1SG	also
#	other	0.1.h:S/v:pred	
fe	‘I grew up without a father.’		

If we look at the annotation of sentence (3), we can see that the first word is not specified as far as the thematic role or the syntactic function, although they could be. The second word contains syntactic annotation. As has been explained in section 2.1.1, the first element indicates the form of the referent. In the present case it refers to the first person, which is not overtly expressed in the sentence. The second element describes the inherent property of the referent, while the last element describes a syntactic function. Thus, the sentence has a covert first person subject. The second syntactic annotation follows after the slash.

Syntactic roles are annotated similarly to thematic roles, following the principle used in GRAID, according to which annotations have the form such as <form.animacy:function>. Table 5 summarizes the annotation options of the subject function.

Table 5

The annotation of the subject

<table>
<tr><th>Abbrev.</th><th colspan="2">Form of referent</th><th colspan="2">Inherent properties of referent</th><th colspan="2">Semantically specified individual form</th><th colspan="2">Function</th></tr>
<tr><td>pro.h:S</td><td>Full pronoun</td><td>Pro</td><td></td><td></td><td rowspan="5">h</td><td rowspan="5">Human</td><td rowspan="6">S</td><td rowspan="6">Subject</td></tr>
<tr><td>0.1.h:S</td><td>Deleted</td><td>0</td><td>First person</td><td>1</td></tr>
<tr><td>0.2.h:S</td><td colspan="3">Deleted
0
Second person</td><td>2</td></tr>
<tr><td>0.3.h:S</td><td>Deleted</td><td>0</td><td>Third person</td><td>3</td></tr>
<tr><td>np.h:S</td><td>Noun phrase</td><td>np</td><td></td><td></td></tr>
<tr><td>pro:S</td><td>Full pronoun</td><td>pro</td><td></td><td></td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td>np:S</td><td>Noun phrase</td><td>np</td><td></td><td></td></tr>
</table>

At present we cannot completely carry out the annotation of every syntactic structure. We cannot, for instance, annotate participial structures that function as clauses and their possible complements, or structures with gerunds etc.:

(4) ChND_061101_TwoTents_flkd.007

mb	*basu-čə-bünü-ndi*	*maa-gəlʲčə*	*ńi-gə-tɨ-gəj*	*kotə-ʔ*
ge	hunt-EMPH-COND.FUT-3DU	what-EMPH.[ACC]	NEG-ITER-PRS-3DU	kill-CNG
SeR		pro:P	0.3.h:S aux.neg	v:pred
fe	‘During hunting they did not kill anything.’			

Non-finite complements (such as, for instance, ‘he went out to hunt’ or ‘she went out to visit with people’) have np:G (goal) as their notation, although this does not completely agree with the definition of the thematic role GOAL.

(5) ChND_061101_TwoTents_flkd.009

mb	*tə*	*kaŋgü-čə-küə-nɨ*	*ŋonəi-ʔ*	*bii-ʔi‿ai-ndi*	*basu-ďa*
ge	well	when-EMPH-EMPH-LOCADV	one.more-ADV	go.away-PF.R-3DU.R	hunt-INF
SeR		adv: time		0.3.h:S v:pred	np:G
fe	‘Once they were going to hunt again.’				

4. Annotation of information structure

Information structure can be conceived of in various ways and several layers of it can be differentiated. One of them is the *theme* vs. *rheme* dichotomy (cf. Holliday, 1967), which approaches the issue from the perspective of the listener, distinguishing between what is new information for the listener and what is known. The terms *topic* vs. *comment* are also used to describe the same thing (cf., for instance, Bloomfield, 1935; Gundel, 1978; Reinhart, 1982), with recent works using the latter terminology. Another layer is that of cognitive representation, where the main focus is the status of information, that is, whether the given information is new for the listener in the given stretch of discourse or not, using the *given* vs. *new* dichotomy (cf. Chafe, 1976; Allerton, 1978). The third layer focuses on the speaker’s intention and operates with the categories of *focus* vs. *background*. The information relevant from the point of view of the speaker will be the focus of the sentence (which is an emphasized constituent of the comment unit), while the information which is less relevant from the point of view of the speaker is the background (cf. Holliday, 1967; cf. *topic* and *focus* in Lambrecht, 1994).

As various authors have pointed out before (cf. von Heusinger, 1999; Büring, 2005, etc.), prosody also plays an important role in structuring information conveyed by sentences. Even though the Nganasan corpus under construction would make it possible to annotate prosodic features, since audio files with the data are also available, but Nganasan information structure has been studied to such a limited extent that annotating prosody would be much beyond the scope of the current project[6]. As a result, at the present stage of the project we concentrate on annotating only the status of the information.

In annotating information structure in our corpus, we follow the annotation guidelines presented in Götze et al. (2007). Here, we will apply only the Core Annotation Scheme including the annotation layers ‘Information Status’ (with the corresponding tags ‘given’, ‘accessible’, and ‘new’). In the project that serves as a model in Götze’s work, further layers such as ‘Topic’ (with the corresponding tags ‘aboutness topic’ and ‘frame setting topic’), and ‘Focus’ (with the corresponding tags ‘new information-focus’ and ‘contrastive focus’) are used (cf. Götze et al., 2007: 148). However, we do not use these at the present stage of our project but may experimentally annotate Topic and Focus in a part of our corpus at a later stage. Similarly, we may annotate information status according to an extended annotation field in some cases (see below).

[6] For an example of annotating the prosody in a spoken language corpus, see e. g. Baumann (2006).

The notions mentioned above (topic vs. comment, new vs. given, theme vs. rheme) do not cover information structure in exactly the same way, but there are some parallels that can be identified between them. (For a summary of this, see, for instance, von Heusinger, 1999 or Zerbian, 2006.)

In any case, in our project we operate with the notions given vs. new, and this means that in a substantial subset of the cases what is annotated as 'new information' will be the same elements that can also be annotated as 'focus' later. The following English language example illustrates this point.

(6) What does Mary eat?
Mary eats *fish.*
background focus
given new

There is no straightforward parallel like this in the annotation of the topic, the reason for which is that new information is part of the unit 'comment'. If we divide the sentence in (6) into topic vs. comment units, we identify *Mary* as the topic, while the rest of the sentence is the comment. If we compare the information structure from the points of view of topic, focus and information status, we get the following pattern:

(7) What does Mary eat?
Mary *eats* *fish.*
topic comment
background Focus
given New

Now we demonstrate the principles of annotating information status. In this case the focus of the examination is what role the information plays in the discourse. In this annotation scheme three notions are crucial: given, accessible, and new.

Given: an entity is given if it has previously occurred in the discourse. This previous occurrence does not necessarily have to be in the immediately preceding sentence but can be a few sentences earlier and being activated again now.

In the extended annotation scheme it is possible to differentiate between referents that are active vs. not active. A referent is active if it occurred in the previous sentence, while it is inactive if earlier than that.

Accesible: a referent is accessible if it has not been mentioned before but can be identified, for instance, from the context of the situation, general knowledge, or the course the discourse takes subsequently. According to Götze's system (2007: 157–160) it is possible to annotate exactly what is known. We do not go into details such as this and use core annotation instead.

New: an element is new in a sentence if it conveys new information in the sentence.

Table 6 below summarizes the abbreviations used for annotating Nganasan information status.

Table 6

The annotation of information status

Information status	**Given**	**Accessible**	**New**
Annotation	giv (underspecified) giv-active giv-inactive	accs (underspecified)	new

5. Conclusion

The summary of the annotation system used in our Nganasan corpus provides an example of multitier annotation which can be extended with further information at any future time. This annotation system makes it possible to do complex searches, searching for various types of information at the same time, which, in turn, can yield insight into interrelationships in the data which previous corpuses

were not able to uncover, such as interrelations between thematic roles, syntactic roles, and morphological form. The following examples

(8) ChND_061101_TwoTents_flkd.001

st	**Ситі маʔ.**	
ts	Sʲiti maʔ.	
tx	*Sʲiti*	*maʔ*
mb	sʲiti	maʔ
mp	sʲiti	mat
gr	два.[NOM]	чум.[NOM]
ge	two.[NOM]	tent.[NOM]
ps	num-n.case	n-n.case
#	other:attr	np
IST	new	
fr	Два чума.	
fe	There are two tents.	

(9) ChND_061101_TwoTents_flkd.002

st	**Сиӡи матәны нилытыгәй ситі нумәәгәй.**				
ts	Sʲiði matəni n’ilitigəj s’iti numəəgəj.				
tx	*Sʲiði*	*matəni*	*ńilitigəj*	*sʲiti*	*numəəgəj.*
mb	sʲiði	ma-təni	ńili-ti-gəj	sʲiti	numə-ə-gəj
mp	sʲiti	maʔ-ntənu	n’ili-ntu-kəj	sʲiti	numə-ə-kəj
gr	два.[GEN]	чум-LOC	жить-PRS-3DU.S	два	парень-ADJ-NOM.DU
ge	two.[GEN]	tent-LOC	live-PRS-3DU.S	two	young.man-ADJ-NOM.DU
ps	num-n.case	n-n.case	v-v.tense-v.pn	num	n.-deriv.adj-n.case.number
#	other:attr	np:L	v:pred	other:attr	np.h:S
IST	give-active			New	
fr	В двух чумах живут два парня.				
fe	In these two tents live two young men.				

Abbreviations

ACC – accusative;
ADJ – adjective;
ADV – adverbial suffix;
CNG – connegative;
COND – conditional;
DERIV – derivational suffix;
DST – destinative;
DU – dual;
EMPH – emphatic element;
FUT – future;
INF – infinitive;
ITER – iterative;
LOCADV – locative adverbial suffix;
N – noun;
NEG – negative;
NOM – nominative;
NUM – numeral;
PF – perfect;
PN – personal suffix;
PRIV – privative;
PRS – present;
PST – past;
R – reflexive;
SG – singular;
TR – transitive;
V – verb.

References

Allerton, D. J. 1978. The notion of 'givenness' and its relation to presupposition and theme. *Lingua* 44. 133–168.

Baumann, Stephan. 2006. Information Structure and Prosody: Linguistic Categories for Spoken Language Annotation. In Sudhoff, Stefan, Denisa Lenertová, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter & Johannes Schließer (eds.). *Methods in Empirical Prosody Research*. Berlin, New York: De Gruyter (= Language, Context, and Cognition 3). 153–180.

Butt, Miriam. 2005. *Theories of Case*. Cambridge: Cambridge University Press.

Büring, Daniel. 2005. *Intonation und Informationsstruktur*. Available online at http://semanticsarchive.net/Archive/jI0OTk3O/buring.ids2005.pdf (accessed 17.03.2015).

Dowty, David. 1989. On the Semantic Content of the Notion "Thematic Role". In *Properties, Types, and Meanings*. Vol. 2. Ed. by G. Chierchia, B. H. Partee, and R. Turner. Dordrecht: Kluwer. 69–130.

Dowty, David. 1991. Thematic Proto-Roles and Argument Selection. *Language* 67. 547–619.

Fillmore, Charles. 1968. The Case for Case. In *Universals in Linguistic Theory*. Ed. by Emmon Bach and Robert T. Harms. New York: Holt, Rinehart & Winston.

Gawron, Jean Mark. 2007. *Aspect, Roles, and Lexical Semantics*. Available online at http://www-rohan.sdsu.edu/~gawron/semantics/course_core/lectures/roles.pdf (accessed 17.03.2015).

Götze, Michael et al. 2007. Information structure. In Dipper, S., Götze, M. and S. Skopeteas (eds.): *Information Structure in Cross-Linguistic Corpora*. Interdisciplinary Studies on Information Structure 07 (2007): 147–187. Available online at http://edoc.hu-berlin.de/oa/reports/reQ5PntJcwYs/PDF/23TFAo8H6FW2.pdf (accessed 05.02.2013).

Grimshaw, Jane. 1990. *Argument Structure*. MIT Press, Cambridge, Massachusetts.

Gundel, Jeanette. 1978. *Stress, pronominalisation and the given-new distinction*. University of Hawaii Working Papers in Linguistics 10/2. 1–13.

Haig, Geoffrey and Stefan Schnell. 2011. *Annotations using GRAID (Grammatical relations and animacy in discourse)*. Introduction and guidelines for annotators. Version 6.0. Available online at http://www.linguistik.uni-kiel.de/GRAID_manual6.0_08sept.pdf (accessed 05.02.2013).

Haig, Geoffrey, Stefan Schnell. 2014. *Annotations using GRAID.* Manual Version 7.0.

Halliday, M.A.K. 1967. Notes on transitivity and theme in English, Part 2. *Journal of Linguistics* 3. 199–244.

Von Heusinger, Klaus. 1999. *Intonation and Information Structure. Habilitationsschrift*. Available online at http://gerlin.phil-fak.uni-koeln.de/kvh/pub/pub99–98/Heusinger1999_Intonation.pdf (accessed 25.02.2014).

Von Heusinger, Klaus. 2002. Information structure and the partition of sentence meaning. In Hajičová, Eva, Petr Sgall, Jirí Hana and Tomáš Hoskovec (eds.), *Prague Linguistic Circle Papers*: Travaux du cercle linguistique de Prague nouvelle série. Vol. 4. 275–305, Amsterdam: Benjamins.

Kittilä, Seppo. 2005. Recipient prominence vs. beneficiary prominence. *Linguistic Typology* 9. 269–297.

Malchukov, Andrej, Martin Haspelmath and Bernard Comrie, eds. 2010. *Studies in ditransitive constructions*: a comparative handbook. Berlin: De Gruyter Mouton.

Schmidt, Thomas, Wörner, Kai. 2005. Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. *Gesprächsforschung* 6. 171–195.

Wörner, Kai. 2010. *Werkzeuge zur flachen Annotation von Transkriptionen gesprochener Sprache*. PhD Thesis, Universität Bielefeld. Available online at http://bieson.ub.uni-bielefeld.de/volltexte/2010/1669/ (accessed 05.02.2013).

Zerbian, Sabine. 2006. Expression of information structure in the Bantu language Northern Sotho. *ZAS Papers in Linguistics* 45. Berlin: ZAS.

Wagner-Nagy B.
Institut für Finnougristik/Uralistik.
Von-Melle-Park, 6, 20146 Hamburg, Deutschland.
E-mail: beata.wagner-nagy@uni-hamburg.de

Szeverényi S.
Institut für Finnougristik/Uralistik.
Von-Melle-Park, 6, 20146 Hamburg, Deutschland.
E-mail: sandor.szeverenyi@uni-hamburg.de

Материал поступил в редакцию 25 апреля 2015.

Вагнер-Надь Б., Северени Ш.

АННОТИРОВАННЫЙ ЗВУКОВОЙ КОРПУС НГАНАСАНСКОГО ЯЗЫКА

Описываются некоторые ключевые моменты разметки, используемой в проекте «Аннотированный звуковой корпус нганасанского языка». Данные в проекте обрабатываются и хранятся в формате EXMARaLDA. Разметка базы данных включает грамматические и частеречные глоссы (созданные в программах Toolbox или Flex), переводы на русский и английский языки; однако настоящая статья посвящена в первую очередь информации о синтаксических ролях, тема-рематических характеристиках и информационной структуре. Для этого используется формат, разработанный другими исследователями и приспособленный авторами для нганасанского языка. Представлена система разметки (теги, термины и их объяснения), проиллюстрированная большим количеством нганасанских примеров.

Ключевые слова: *нганасаны, аннотация.*

Вагнер-Надь Б., доктор, профессор.
Институт финно-угристики и уралистики. Университет Гамбурга.
Von-Melle-Park, 6, 20146 Hamburg, Deutschland.
E-mail: beata.wagner-nagy@uni-hamburg.de

Северени Ш., доктор.
Институт финно-угристики и уралистики. Университет Гамбурга.
Von-Melle-Park, 6, 20146 Hamburg, Deutschland.
E-mail: sandor.szeverenyi@uni-hamburg.de