

naomi.nagy@utoronto.ca

m.meyerhoff@auckland.ac.nz

naomi.nagy@utoronto.ca

Extending ELAN into Quantitative Sociolinguistics

Abstract

ELAN (tla.mpi.nl/tools/tla-tools/elan, Wittenburg et al. 2006) has established itself as a valuable tool for language documentation and is frequently used for transcription and multi-tier mark-up illustrating levels of linguistic structure as well as translations and glosses. We illustrate an extension to its utility: **extracting and coding tokens of linguistic variables for quantitative analysis in the variationist sociolinguistic framework**. There are a number of benefits of this approach that can improve the way sociolinguistic research is conducted and which relatively easily extends the scope of language documentation and conservation work. Since ELAN is well-known in the LDC community, we focus on how it can facilitate the use of our corpora for the study of synchronic variation:

- seamless connections between recording, transcript, and coding of the dependent variable (response) and independent variables (predictors). This facilitates revision and intercoder reliability tests.
- exportability to Excel, R, Rbrul, Goldvarb, SPSS, ...
- importability of transcripts from Word/text files
- complex searches and concordancing capabilities to speed up token extraction
- archivability of all mark-up related to each data file in a consistent and small-file-size format

Endangered language documentation today is dealing with synchronic variation within the speech community in greater detail than it has in the past (*cf.* Flores Farfán & Ramallo 2010). There are principled reasons for this. Grenoble argues that “More research is pressingly needed in the area of contact-induced change and language attrition,” (2010: 67) and that “linguists can help educate speakers in dialect awareness, to understand that variation is the natural result of language change and is found in vital languages which are robustly spoken” (Grenoble 2010:83).

This presentation reviews the major arguments justifying this approach and provide a how-to demonstration, illustrating ELAN’s functionality for variationist research with ongoing work on variation in subject pronoun presence in Faetar and N’kep, endangered languages spoken in southern Italy and Vanuatu (respectively).

Why look at variation?

Practical

- Speakers, especially of endangered languages worry about variation as an indication of loss
- “Linguists can help educate speakers in dialect awareness, to understand that variation is the natural result of language change and is found in vital languages which are robustly spoken”

(Grenoble 2010:83).

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

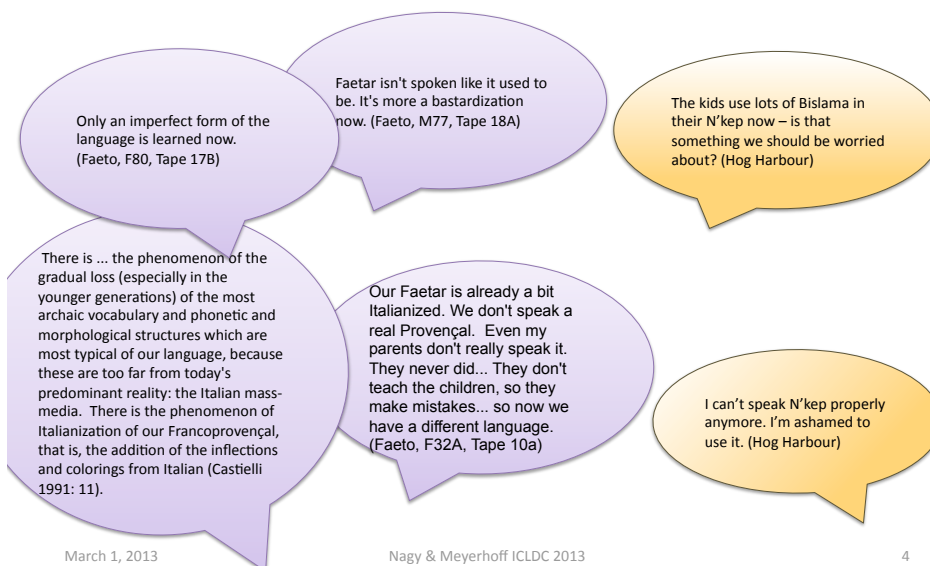
Theoretical

- Better understand structure and behaviour of language
- “More research is pressingly needed in the area of contact-induced change and language attrition.”

(Grenoble: 2010: 67)

3

Why look at variation?



March 1, 2013

Nagy & Meyerhoff ICLDC 2013

4

Imperatives for studying variation in minority and endangered languages

- **theoretical:** linguists acknowledge variation and its role in language change
- **practical:** it's what communities and we become aware of very quickly
- **moral/social:** it's what our speakers are interested in

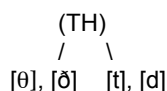
How does this mesh with *your* experiences in language documentation?

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

5

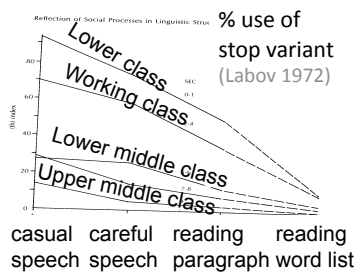
- Variation exists in all languages.
- It is systematic and rule-governed.



- in a big language (English)
- phonetic
- NOT indicative of change in progress



- in a small language (**Faetar**)
- morphosyntactic
- indicative of change in progress?



- in a small language (**N'kep**)
- morphological
- status unclear

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

6

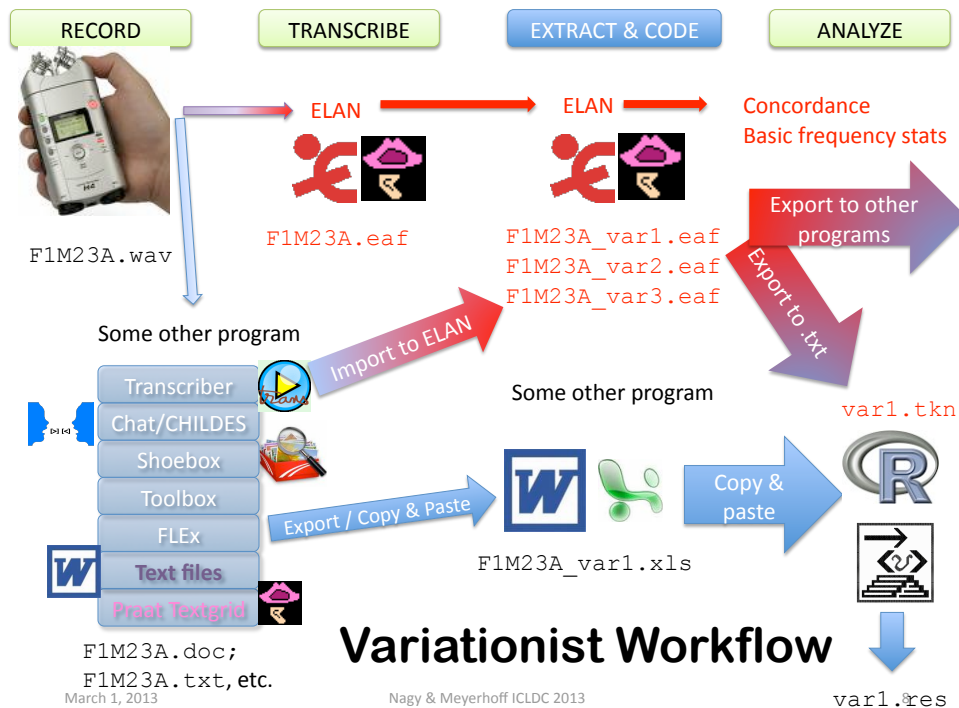
Two purposes of studying variation

- *descriptive*: in which contexts is it normal/typical for speakers to use which forms?
 - what are the rules for the variation?
- *heuristic*: what can we learn about this language from the variation?
 - what is the best analysis of this structure/form?

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

7



EXAMPLE FROM FAETAR (PRO-DROP)



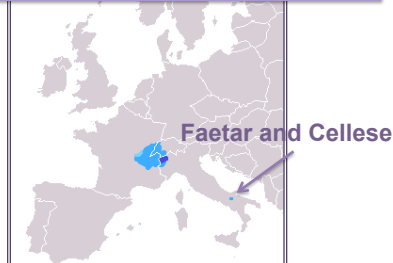
The variable
(pro-drop)
/ \
[i] ∅

Language Franco-Provençal [FRP]
Population 70,000 in Italy (1971 census);
~600 in Faetar dialect (Nagy 2000).
Language use Most domains. Also use Italian
or Piemontese [pms].
(www.Ethnologue.ca)

Examples of alternation

F11B: Allorə i ʌ est a kaze
So, he is at home. (F11B.1.tr:1)

F11B: Allor ike ∅ sə truwund ingjen la vi
So, here he is in the street. (F11B.3.tr:2)



March 1, 2013



Nagy & Meyerhoff ICLDC 2013



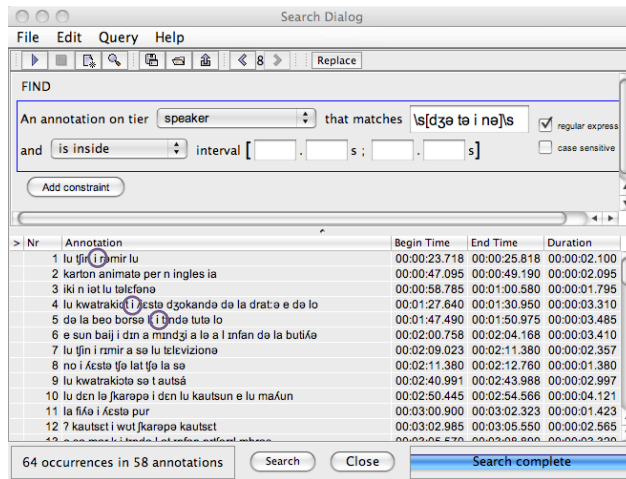
9

Why study variation in Faetar pronouns?

- Remember Grenoble's encouragement...
- Do the patterns of pro-drop in Faetar suggest the effect of language contact (with Italian)?
 - Compare the distribution of *strong*, *weak* and *null pronouns*.
 - Compare the distribution among different age groups.

Faetar Search Results

Regular expression search for weak-form subject pronouns:



Weak Forms		
	Sing.	Plural
1 st	dʒə	nə
2 nd	tə	və
3 rd	i	i

Strong Forms		
	Sing.	Plural
1 st	dʒi	nu(s)
2 nd	ti	vu(s)
3 rd	i(ʌ)	is

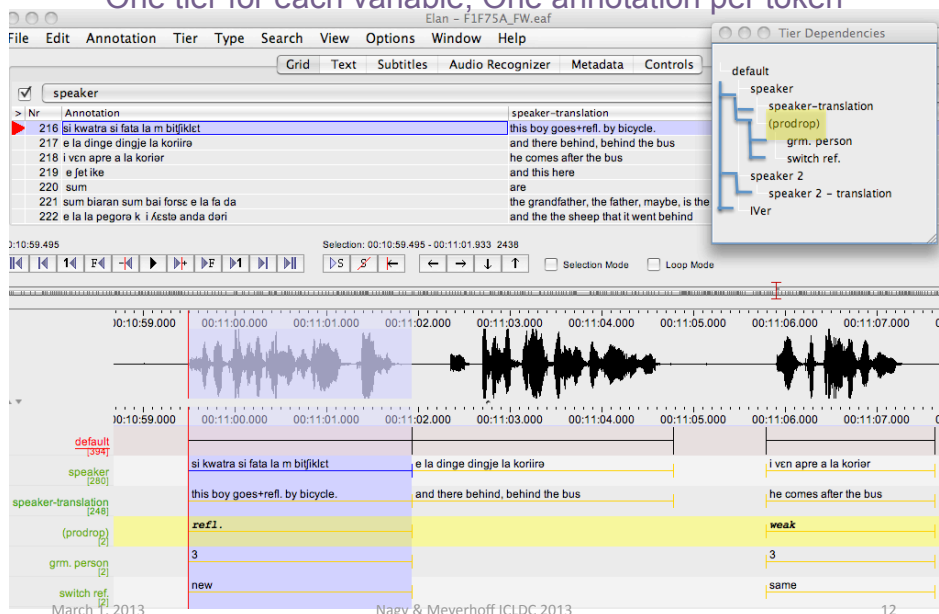
Null Form		
	Sing.	Plural
1 st	∅	∅
2 nd	∅	∅
3 rd	∅	∅

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

11

Coding & Extracting: One tier for each variable; One annotation per token



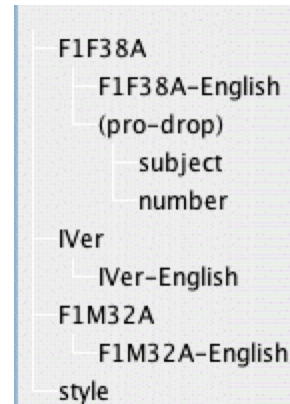
March 1, 2013

Nagy & Meyerhoff ICLDC 2013

12

Coding & Extracting

- Linguistic and stylistic factors are coded directly in ELAN, each on their own tier.
- Advantages:**
 - See all the context you need, and hear it, check it in Praat, as you code each factor.
 - From ELAN, create a .txt file for multivariate analysis (using Rbrul or Goldvarb or ...)
 - Can (repeatedly) revise codes in ELAN and quickly recreate the data file.



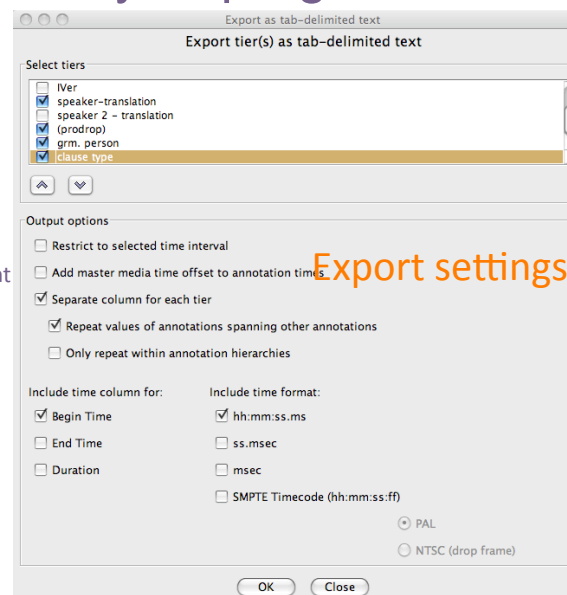
March 1, 2013

Nagy & Meyerhoff ICLDC 2013

13

Export to analysis program

- File > Export as... > Tab delimited Text. Make sure the filename specifies the speaker and ends in ".txt"
- Select all the tiers that have relevant labels or transcriptions in them.
- Select: Separate column for each tier
- Save as a .txt file.
- Open the .txt file in Excel (use Import, skip directly to "Finish.")



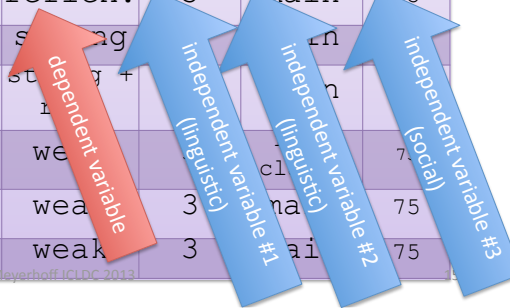
March 1, 2013

Nagy & Meyerhoff ICLDC 2013

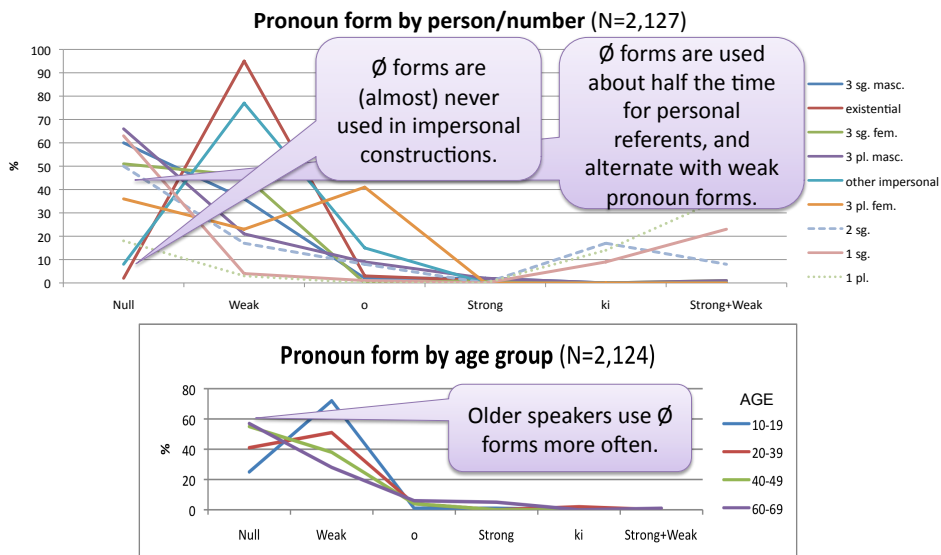
14

(Sorted) Exported Coding File

Begin Time	Extract	Translation	(prodrop)	grm. person	clause type	Speaker age
13:18.4	anjat la at:	there is the cat	0	expl	main	75
13:30.5	o set i ʌestə kə sə fət fə dəʃkʌnd pə dəso la la lu tawula	expl. this it is that refl. makes hide under the the table	both	3	main	75
11:38.8	kə s andə fərmá kə nə tʀavərsə: la vi	that has stopped +refl. so-that they can cross the street	refl	4	rel. clause	75
10:59.5	si kwatra si fata la m bitʃiklet	this boy goes+refl. by bicycle.	reflex.	3	main	75
11:36.6	set e lu stopə	This is the stop-sign	st ag +			
13:37.4	set ike sə vatə sə rir dəso lu dʒəri:ɛ	this here refl. (?) under the (?)	st ag +			
11:14.2	e la la pegorə k i ʌestə anda dəri	and the the sheep that it went behind	we			75
12:35.7	i tndə lo dʒəla:tə pə tu:t i tndə	he holds the icecream for everyone	wea	3	ma	75
12:37.8	i tndə	he holds	wea	3	ai	75



Faetar variation patterns

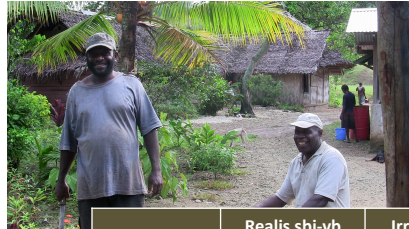


March 1, 2013

Nagy & Meyerhoff ICLDC 2013

16

EXAMPLE FROM N'KEP: [tə-, təm-] PREFIX



Hog Harbour



	Realis sbj-vb	Irrealis sbj-vb
1 singular	nam-; m-	nac-
2 singular	nem-; nəm-	nec-
3 singular	m-	cə-
1 pl exclusive	cam-; t(ə)-	ca-; t(ə)-
1 pl inclusive	t(ə)-	t(ə)-
2 plural	cam-; t(ə)-	cə-
3 plural	cam-; cəm-	ca-
Indefinite	təm-; təm-	te-; tə-

tə- seems to be more than 'indefinite' – or is it spreading?



March 1, 2013

Nagy & Meyerhoff ICLDC 2013

17

Why study variation in N'kep subjects?

- Unclear what the [tə-] and [təm-] prefixes on verbs mark
 - person and number?
 - subject properties (e.g. generic, indefinite subject)?
 - event properties (e.g. backgrounded, subordinate event)?
 - clause aspect?

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

18

Independent variables coded

- Person and number of the subject referent
 - 1s, 2s, 3s, 1p.incl, 1p.excl, 2p, 3p
 - (in)definiteness of subject
- Same-subject versus switch-subject
 - same, switch, new
 - two models for subset chains
- Event type
 - future, other irrealis (e.g. desired), negative, imperfective (? : *L-verb*), inflected 'when', *go-come* as V2, elsewhere (realis event/state)
- Generation
 - older, middle, younger

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

19

N'kep coding and tier organisation

The screenshot shows the ELAN software interface with a transcription and its coding. The transcription is in the top panel, and the coding is in the bottom panel. The coding is organized into tiers, including transcription, translation-Bislama, English translation, subject type, same/switch subject, sbj pers&no, event semantics, same/switch subject-cp, (in)definiteness of sbj, and Notes and comments. A circled area highlights the subject type and same/switch subject-cp tiers.

Tier	Code	Value
subject type	undefreal	defire
same/switch subject	switch	switch
sbj pers&no	5	3
event semantics	REAL-event	REAL-
same/switch subject-cp	switch	switch
(in)definiteness of sbj	participant	3-nam

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

20

Extracting data for analysis

Search Dialog

File Edit Query Help

Replace

FIND

An annotation on tier "subject type" that matches string indef:irreal

Speaker uses [tə-] for a range of persons/numbers

Other factors can be treated as independent (predictor) variables

Nr	Annotation	Child
1	indef:irreal	[switch] [6] [REAL-event] [switch] [generic]
2	indef:irreal	[switch] [6] [WHEN] [switch] [generic]
3	indef:irreal	[switch] [6] [REAL-event] [switch] [generic]
4	indef:irreal	[same] [5] [REAL-event] [same] [participant]
5	indef:irreal	[same] [5] [REAL-event] [same] [participant]
6	indef:irreal	[same] [5] [REAL-event] [same] [participant]
7	indef:irreal	[same] [5] [REAL-event] [same] [participant]
8	indef:irreal	[switch] [4e] [REAL-event] [switch] [generic]
9	indef:irreal	[same] [4e] [REAL-event] [same] [generic]

9 occurrences in 9 annotations

Search Close Search complete

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

21

Results for tə-, təm- distribution

Referent type	Weighting	%	N
1st and 2nd person	0.91	58	162
Generic, 3rd person indefinite or non-specific, situation or weather	0.27	14	164
3rd person definite or specific	0.22	10	483
Event			
FUTURE-event	< 0.001	0	27
IRREALIS-event	0.85	15	41
NEG-event	0.95	8	41
L-verb	0.99	18	84
REALIS-event	0.99	21	509
go-come	0.99	28	76
WHEN	> 0.99	40	30
Same vs switch reference			
Same referent	0.56	23	152
New	0.54	16	84
Switch referent, incl. proper subset of previous sbj	0.40	17	293

Doesn't look like it marks prototypically 'indefinite' subjects

Maybe it has something to do with discourse topicality?

Generation of speaker (older, middle, younger) is not a significant effect.

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

22

Revisiting Grenoble

- Language variation occurs in vital, as well as endangered languages
 - variation *per se* is not a ‘problem’
 - but the community may perceive it to be.

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

23

Create spiffy WWW examples



Customizable Presentation of ELAN Documents

Users' Manual

Draft · February 2010

For CuPED version 0.3.14

<http://sweet.artsrn.ualberta.ca/cdcox/cuped/>

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

24



CuPED turns an ELAN .eaf into animated .html

okruzheny anglijskim.
How were you taught, what measures did they (parents) take to teach you Russian? It's difficult when you're surrounded by English.

Ну когда, когда были маленькие, конечно дома старались говорить по-русски.
Nu kogda, kogda byli malen'kie, konechno doma staralis' govorit' po-russki.
Well, when we were little, we of course tried to speak Russian at home.

Папа, ему тяжелее было говорить по-английски, мама лучше говорила по-английски.
Papa, emu tjazhelee bylo govorit' po-anglijski, mama luchshe govorila po-anglijski.
Dad, he had a harder time speaking English, mom was better at speaking English.

И я даже всегда должна была, как сказать, переводить иногда,
I ja dazhe vseгда dolzhna byla, kak skazat', perevodit' inogda,
And I always had to, how to put it, translate sometimes,

или как-то нужно было что-нибудь спросить, или телефини-- телефинировать кому-то, я должна была всегда это.
ili kak-to nuzhno bylo chto-nibud' sprositi', ili telefini- telefinirovat' komu-to, ja dolzhna byla vseгда jeto,
or if a question needed to be asked, or to phone someone, I always had to do it.

Speaker: R2F53A
Language: Russian
Generation: 2nd
Age: 53
PLAY SAMPLE

<http://sweet.artsrn.ualberta.ca/cdcox/cuped>

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

25

References

- Flores Farfán, J. & F. Ramallo. 2010. *New Perspectives on Endangered Languages*. Amsterdam/Philadelphia: Benjamins.
- Grenoble, L. 2010. Language vitality and revitalization in the Arctic. In J. Flores Farfán & F. Ramallo. *New Perspectives on Endangered Languages*. Amsterdam/Philadelphia: Benjamins. 65-91.
- Labov, W. *Sociolinguistic Patterns*. Philadelphia: University of PA Press.
- Meyerhoff, M. 2010-2013 Documentation of N'kep (north Vanuatu): Structure and variation. <http://www.hrelp.org/grants/projects/index.php?projid=254>.
- Nagy, N. 2000. *Faetar*. München: Lincom Europa.
- Nagy, N. 2009. Heritage Language Variation and Change in Toronto. <http://projects.chass.utoronto.ca/ngn/HLVC>.
- Nagy, N. 2011. *Lexical Change and Language Contact: Faetar in Italy and Canada*. *Journal of Sociolinguistics* 15:366-382.
- CuPED <http://sweet.artsrn.ualberta.ca/cdcox/cuped>
- Ethnologue www.ethnologue.org
- Praat <http://www.fon.hum.uva.nl/praat/>
- Goldvarb <http://individual.utoronto.ca/tagliamonte/goldvarb.htm>
- Rbrul <http://www.danielezrajohnson.com/rbrul.html>
- R <http://www.r-project.org>

March 1, 2013

Nagy & Meyerhoff ICLDC 2013

26