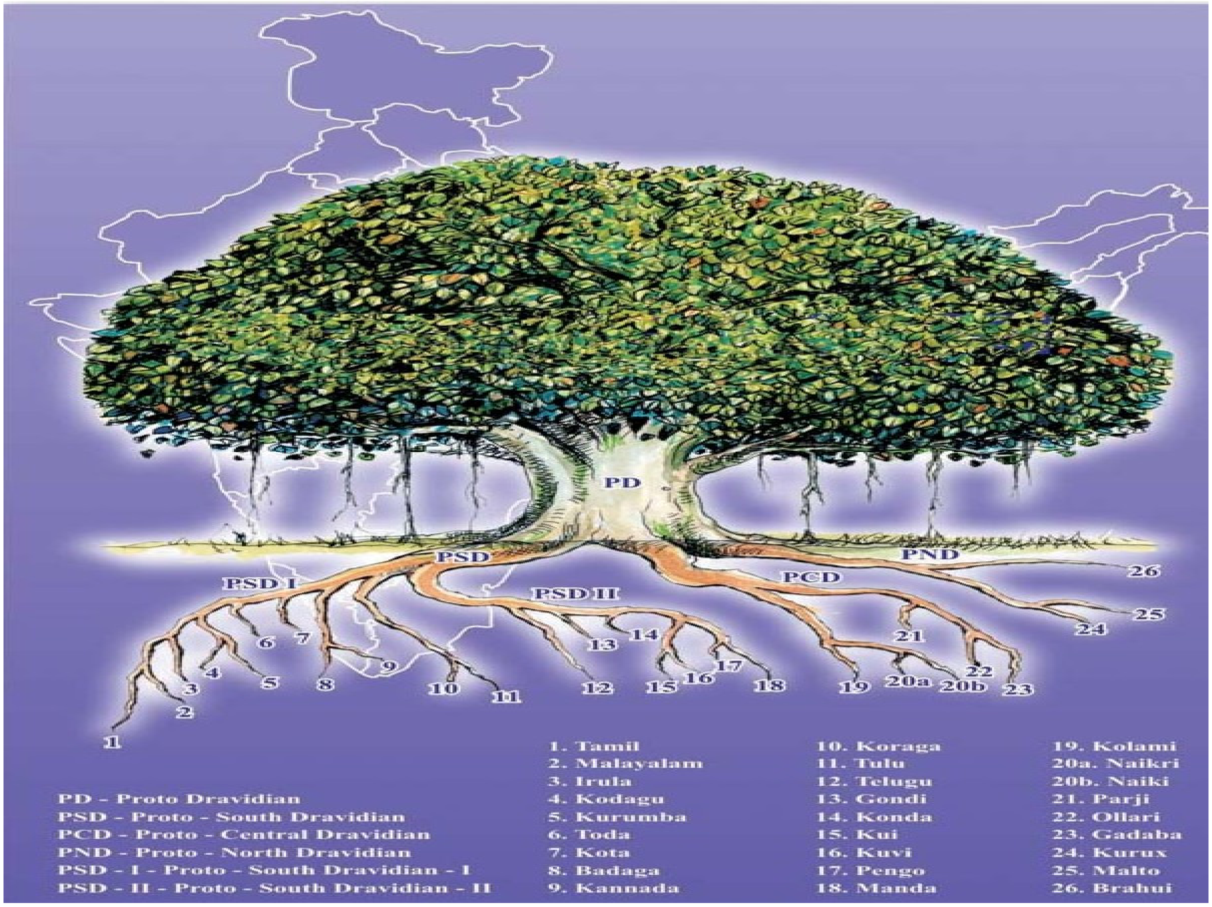
National Workshop on
Standardisation of IT Enabled Transliteration, Glossing and Meta-language for Dravidian Languages

Draft of the Standards of IT enabled Transliteration, Glossing and Meta-language for Dravidian Languages circulated in advance among the participants of the workshop, 6-10 March 2017



Organised by

CPeDL

Centre for Preservation of Endangered Dravidian Languages

Funded by University Grants Commission

Dept. of Dravidian & Computational Linguistics

Dravidian University, Kuppam AP

Content

Organisers

Dr. M. C. Kesava Murty
Assistant Professor, Deputy Coordinator SAP-DRS
Coordinator of the workshop

P. Sreekumar
Assistant Professor, Deputy Director of CPEDL
Deputy Coordinator of the workshop

Prof. G. Balasubramanian
Rector, Dravidian University; Director of CPeDL & Coordinator of SAP-DRS

Dr. Ganeshan Ambedkar
Associate Professor, Head of the Department

Dr. M. Prasad Naik
Assistant Professor

Draft compiled by
Mr. Bharath Kumar M S (Research Scholar)
Mr. M Rajakrishna (Research Scholar)
Mr. K. Suryanarayana (Research Scholar)
Mr. D Raja Rao (Research Scholar)

1. Introduction: Linguistic studies are much advanced scientific practice today. Linguistic literature is addressing global linguistic community irrespective of the paradigm, family and the size of the languages. In the midst of the most alarmed situation of language endangerment, linguistic data and other language resources are considered as one of the common goods to humanity. Therefore, preservation and presentation of language data with maximum linguistic information is a prime concern today. Thus, standardization of **transliteration**, **glossing**[1] and **metalanguage** which we follow are prerequisites for the smooth communication and easy access of linguistic literature at global level. There are number of standards proposed by Lehamann (1982:199-224), Sebastian (2002), Bird Steven & Gary Simons (2003) and Haspelmath and Cormie (2008) etc. Modern linguistic studies in India are more advanced in four language families. Even though, no standard is consistently and widely following in transliteration, glossing techniques and meta-language. This problem is not at all addressed in the linguistic studies of Dravidian family of languages. Standardisation of transliteration, glossing and meta-language standards for Dravidian languages is an imperative today. In this context, this workshop is proposed. The outcome of the workshop shall be further published through *International Journal of Dravidian Linguistics* (IJDL) for wider discussion.

2. Objectives of the workshop: To review the existing status of transliteration, glossing and meta-languages schemes used in Dravidian languages.

1. To develop a standard transliteration, transcription system for each major Dravidian language by the linguists of each language.
2. To develop a glossing standard for Dravidian languages.
3. To develop a standard meta-linguistic representation and abbreviation for Dravidian languages.

[1] **Glossing** is an analytical technique and a way of presentation of linguistic data. In this context glossing is defined as an analytical description and a technique of annotation of linguistic data into different levels of representation with meaning and grammatical category labels.

4. To address the issues of Unicode fonts in the communication of Dravidian language data.
5. To develop pedagogic means to disseminate the new standards among the linguists.

3. Structure of the workshop: Preparatory phase: The Dept. of Dravidian and Computational Linguistics will prepare a draft standard and sent to the resource persons of each language in advance. Review report of each resource persons shared among the participants.

1. **Workshop stage**: Presentation of the review report by each resource person and discussion. Finalize the transliteration, glossing and meta-language standards and develop pedagogic strategies to disseminate the standards.
2. **Post workshop stage**: Publish the proposal as group outcome in the **International Journal of Dravidian Linguistics (IJDL)** for wider dissemination.

4. Organization of the workshop: 20 linguists across Dravidian languages will participate in the workshop. The standards developed in this workshop can be adapted by International Journal of Dravidian Linguistics as ***Dravidian University Linguistic Standard***.

5. Outcome of the workshop:

1. A standard transliteration scheme for each major Dravidian language including non-literary languages.
2. A common glossing standard for Dravidian languages.
3. A common meta-language and abbreviations for Dravidian languages.
4. Technological standard for composing Dravidian language data.
5. A pedagogic guideline to disseminate the new standard.

A team publication in the *International Journal of Dravidian Linguistics.*

Resource Persons

Arulmozi, S (Dr) Asst. Professor,
Dept. of Centre for Applied Linguistics & Translation Studies, University of Hyderabad,
Hyderabad – 500 046

Basavaraja Kodagunti, (Dr) Asst. Professor,
Dept. of Kannada, Coordinator, Dept of Linguistics,
Central University of Karnataka, Kodugunti,
Kalaburgi – 585 3767. Karanataka.
Mobile No. 09916053057

Chinmay Vijay Dharurkar, Asst. Professor,
Dept. of Linguistics,
School of Comparative Literature,
Central University of Kerala. Kasaragod
Mail: chinmay@cukerala.edu.in
Mobile No. 9371677728

Giridhar Professor (Prof)
1132 Ist Cross, Lalithadri Road,
Kuvempu Nagara, Mysore 570 023.
Mobile No. 09481531391

Gnanasundaram (Prof)
C/o Dr. Ananda Vadivelu, HIG
S. Vanika IIFM, Residential Colony, Kotara,
Sultana Bagh, Near PGT Chowk.
Bhopal – 462 003.
Mobile No. 09449086896

Maheswaran, C (Dr)
Site No 8, Teacher's Colony Ex.
NGO Colony- Via and Post
Coimbatore-641 002

Meti Mallikarjun, (Dr), Associate Professor,
Dept. of Linguistics, Sahyadri Arts College,
Kuvempu University, Vidyanagar – 577 203.
Shivamogga, Karnataka.
Mobile No. 09448871441

Murigeppa, A (Prof)
H No 172 D5, SSF-407, Fourth Phrase,Yelahanka New Town, Bangalore-64

Panicker G, K, (Prof) International School of Dravidian Linguistics, Kerala
Mail: ijdlisdl@gmail.com
Mobile No. 9387828502

Praveen. G (Dr)
Assistant professor,
Dept. of Computational Linguistics: Indian Grammatical Tradition
Banaras Hindu University
Mobile No:91+8179407778

Ramakrishna Reddy,B (Prof)
Gayathri Towers, RTD,
Secunderabad – 500 017

Ramakrishnan A G (Prof)
Professor and Chairman, EE,
Dept. of Electrical Engineering,
Indian Institute of Science,
Banglore-560 012

Ramamoorthy, L (Prof)
CIIL, Hunur Road, Manasagangotri,
Mysore – 570 006.
Mail: ramamoorthyciil@gmail.com
Mobile No. 0821 – 2345020

Ramaswamy, C (Prof)
A 108, Koncept Nest, 6
Bangalore - 560 026

Rangan K, (Prof)
H No:25, Chelleynagar,
Tamil University Post,
Thanjavur, Tamil Naidu-613010

Ravisankar S. Nair, (Prof)
Assoc. Professor,
Dept. of Linguistics,
School of Comparative Literature,
Central University of Kerala.
Mail: ravisankarnair101@gmail.com
Mobile No. 09447375696

Ravisankar, (Prof) Director
Pondicheri Institute of Linguistic and Culture,
LAWSPT Pondicheri 605 008.
Mobile No. 09443187650.

Sobha L, (Dr)
Anna University K B Chandrasekhar Research Centre ,
Chennai

Sreenathan M, (Prof)
Malayalam University,
Tippu Sulthan Road,
Malappuram Dist, Tirur, Vakkad,
Kerala-676502

Subramanya Sharma (Dr)
Guest lecturer
Dept. of Linguistics
Osmania University, Hyderabad

Uma Maheshwar Rao G (Prof)
Centre for Applied Linguistics &Translation Studies,
University of Hyderabad,
Hyderabad

Venkitaswamy, T
Editorial Assistant
Prasaraga, Dravidian University

Venugupala Panikkar G, (Prof)
Vanni, Farook Collage post,
Calicut District-673632

Viswanatha Naidu
Sweden

Viswanatham, (Prof)
102, 4th Main Road, Gokulam 3rd Stage,
Mysore 570 002.
Mobile No. 09480770557.

Schedule of the Workshop (tentative)

Date	Activities					
	10 AM to 11 AM	**11 AM to 1 PM**		**Brake 1PM to 3PM**	**3 PM to 4 PM**	**4 PM to 5 PM**
6.03.2017	Inauguration	Introduction of the workshop (PS)	Presentation of transliteration (Bharath)		Discussion	Discussion
7.03.2017	Presentation of Grammatical categories (MCK)	Discussion			Presentation of abbreviations (Raja)	Discussion
8.03.2017	Presentation of Glossing (PS)	Discussion			Discussion	Discussion
9.03.2017	Presentation of Technology issues (MCK)	Discussion			Group Discussion	Group Discussion
10.03. 2017	Presentations of transliteration (Bharath)	Presentation of Glossing (PS)	Presentation of abbreviation (MCK)		Discussion on dissemination	Valedictory function

General Guidelines

This is only a draft material of the workshop. There are three components in this draft: transliteration of 25 Dravidian languages, a list of abbreviations of grammatical terminologies and glossing standards. A set of questions is attached to each component. We request you kindly go through each component and note your comments for discussion. Following are the general guidelines.

1. Transliteration: We have presented the generally used transliteration of 25 Dravidian languages. You please go through transliteration of each language and respond on each based on the questions we have asked and beyond the questions. If you can propose a set of general rules for standardization of transliteration of all Dravidian language that is also can be discussed.

2. Grammatical categories: We are not proposing any change or new grammatical categories in this document. Generally used grammatical terminologies are presented. However, certain conceptual issues of semi-vowels, retroflex as a point of articulation can be discussed. We request you kindly raise any issues regarding the standardization of grammatical categories in Dravidian languages.

3. Abbreviations: We have presented the abbreviations of grammatical categories, the name of the languages and journals. You please go through it and present your critical comments on it.

4. Linguistic Glossing: We are proposing the Leipzig Glossing Rules jointly developed by Dept. of the Max Planck Institute for Evolutionary Anthropology and Dept. of Linguistics of the University of Leipzig (Bickel, Comrie, and Haspelmath 2004) for Dravidian languages. Based on Malayalam data ten rules of Leipzig Glossing have been presented. We request you kindly go through it and suggest further modifications.

5. Technology issues: We are proposing Arial Unicode MS font for composing text data of Dravidian languages.

Tamil (Annamalai and Steever 1998: 100-28)

Vowels

	Front		Mid		Back	
	Short	Long	Short	Long	Short	Long
High	i	ī			u	ū
Mid	e	ē	Λ		o	ō
Low		(ǣ)	a	ā		

Consonants

	Labial	Dental	Alveolar	Retroflex	Palatal	Velar	Glottal
Stops:							
Voiceless	P	t		ṭ	C	k	
Voiced	(b)	(d)		(ḍ)	(j)	(g)	
Tap		r	[r]				
Nasal	m	n	[n]	ṇ	ñ	ṅ	
Lateral		l		ḷ			
Glide	v				y		

Questions

1. Is this transliteration representing the phonemic system of Tamil without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

2.Malayalam (Asher and Kumari 1997:405-50)

Vowels

i	ī				u	ū
e	ē				o	ō
		(æ)				
			a	ā		

Consonants

	Labial	Dental	Alveolar	Retroflex	Palatal	Velar	Glottal
StopsVoiceless	p	T	t	ṭ	c	k	
Voicelessasp	ph	th		ṭh	ch	kh	
StopsVoiced	b	d		ḍ	j	g	
Voicedasp	bh	dh		ḍh	jh	gh	
Fricative	(f)		s	ṣ	ś		h
Nasal	m	n	n̲	ṇ	ñ	ṅ	
Liquid							
Tap/trill			r, r̲				
Lateral			L	ḷ			
Approx.				ẓ			
Glide	v				y		

Questions

1. Is this transliteration representing the phonemic system of Malayāḷam without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

3.Koḍagu (Balakrishna 1976: 1-4 ff.)

Vowels

i	ï	u
e	ë	o
	a	

Consonants

P b	t d	ṭḍ	cj	kg	
m	n	ṇ	ñ	ṅ	
	s	ṣ	š		h
	l	ḷ			
	r				
v			y		

Questions

1. Is this transliteration representing the phonemic system of Kodagu without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

4.ĀluKuṟumba (Kapp 1987:409)

Vowels

i		ï	u
e		ë	o
	a		

Consonants

p	t		c	ṭ	k
b	d		j	ḍ	g
m	n	.		ṇ	ŋ
		r ṟ			
		l		ḷ	
v		s	y		

Questions

1. Is this transliteration representing the phonemic system of kurumba without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

5.Iruḷa (Zvelebil1973:§1)

Vowels[6]

i		ï				ü		u
	e		ë		ö		o	
				a				

Consonants

p	t	t̠	ṭ	c	k
b	d	d̠	ḍ	j	g
m		n	ṇ		
			ḷ		
		l			
		r			
		r̠	ṛ		
v				y	

Questions

1. Is this transliteration representing the phonemic system of Irula without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

6.Kota (Emeneau1944, KotaTexts,UCPL2:1.5–18)

Vowels

i	ī			u	ū
e	ē			o	ō
		a	ā		

Consonants

P	t	t	ṭ	č	k
b	d	d̲	ḍ	J	g
m	n		ṇ		ṅ
		l	ḷ		
		r			
v				y	

Questions

1. Is this transliteration representing the phonemic system of Kota without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

7.Toda (Emeneau1958:15–66,1984:7,11;Shalevetal.1994:19–56)

Vowels

	Front		Central	Back	
	Unrounded	Rounded	Rounded	Unrounded	Rounded
High	i i:	ü ü:		ï ï:	u u:
Mid	e e:		ö ö:		o o:
Low			a	a:	

Consonants

					Alveolo-		
	Labial	Dental	Post-dental	Alveolar	palatal	Retroflex	Velar
Stopand	p b	t d	c z,	t̠ d̠	č ǰ	ṭ ḍ	k g
Affricate			[ts dz]		[tš dž]		
Nasal	m			n		ṇ	(ŋ)
Fricative	f	*θ*					*χ*
Trill			r	r̠		ṛ	
Lateral				ł̠ l		ł ḷ	
Sibilant			s (z)	s̠ (z̠)	š ž	ṣ ẓ	
Continuant					y		w

Questions

1. Is this transliteration representing the phonemic system of Toda without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

8ModernKannaḍa(Sridhar1990:291–313)

Vowels

	Front	Central	Back
High	i ī		u ū
Mid	e ē		O ō
Lower-mid	æ		
Low		a ā	

Consonants

	Labial	Dental–alveolar	Retroflex	Palatal	Velar–glottal
Stop-vl	p	t	ṭ	c	k
Stop-vd	b	d	ḍ	j	g
Fricative	f	sz	ṣ	ś	h
Nasal	m	n	ṇ		
Lateral		l	ḷ		
Semivowel	v			y	

Questions

1. Is this transliteration representing the phonemic system of Kannada without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

9Baḍaga[9](HockingsandPilot-Raichoor1992:xvi)

Vowels

ī	i			u	ū
ē	e			o	ō
		a	ā		

Consonants

p	t	ṭ	c	k
b	d	ḍ	j	g
			s	(h)
m	n	ṇ		
	r			
	l	ḷ		
v			y	

Questions

1. Is this transliteration representing the phonemic system of Badaga without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

10Tuḷu(D.N.S.Bhat1998)

Vowels

i	ī	ï		u	ū
e	ē			o	ō
ɛ	ɛ̄	a	ā		

Consonants

	Labial	Dental	Retroflex	Palatal	Velar
Stops:					
Voiceless	p	t	ṭ	c	k
Voiced	b	d	ḍ	j	g
Sonorants:					
Nasal	m	n	ṇ	ñ	ṅ
Oral	v			y	
Lateral		l	ḷ		
Trill		r			
Fricative		s			h

Questions

1. Is this transliteration representing the phonemic system of Tulu without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

11Koraga(D.N.S.Bhat1971:4)

Vowels

i	ī	ï		u	ū
e	ē			o	ō
		a	ā		

Consonants

p	t	ṭ	c	k
b	d	ḍ	j	g
m	n			ŋ
v	r		y	
	l			
	s			

Questions

1. Is this transliteration representing the phonemic system of Koraga without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

SouthDravidianII(South-CentralDravidian)

12Telugu(Krishnamurti1998d:260)

Vowels

i	ī			u	ū
e	ē			o	ō
	(ǣ)[10]				
		a	ā		

Consonants[11]

	Labial	Denti-alveolar	Retroflex	Palatal	Velar
Stops:					
Voiceless	p ph	t (th)	ṭ ṭh	c ch	k kh
Voiced	b bh	d dh	ḍ ḍh	j jh	g gh
Fricative	f	S	ṣ	ś	h
Nasal	m	n	ṇ		
Lateral		l	ḷ		
Flap		r			
Semivowel	w				y

Questions

1. Is this transliteration representing the phonemic system of Telugu without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

13Gondi(overallpatternofdifferentdialects)(Rao1987b:101)

Vowels

i	ī			u	ū
e	ē			o	ō
		a	ā		

Consonants

p	b	t	d		ṭ	ḍ	c	j	k	g	
				s							h
				r	ṛ						
				ṟ							
				l	ḷ						
		m		n	ṇ				ŋ		
		w						y			

Questions

1. Is this transliteration representing the phonemic system of Gondi without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

14Koṇḍa/Kūbi(Krishnamurti1969a:185–6)

Vowels

i	ī			u	ū
e	ē			o	ō
		a	ā		

Consonants

Obstruents									
Stop	p b	t d					ṭ ḍ	kg	
Fricative				s	z				(h)
Trill				R	r̠				
Sonorants									
Flap				r			ṛ		
Nasal	m			n			ṇ	ŋ	
Lateral			l	ḷ					
Semiconsonant	w					y			

Questions

1. Is this transliteration representing the phonemic system of Konda without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

15Kui(Winfield1928:1–5)

Vowels

i	ī			u	ū
e	ē			o	ō
		a	ā		

Consonants[12]

p	b	t	d	ṭ	ḍ	s	j	k	g	
		s								h
m		n				n				
		l								
		r				ṛ				
v										

Questions

1. Is this transliteration representing the phonemic system of Kui without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

16Kuvi(Israel1979:§1)

Vowels

i		u
e		o
	a	

Consonants

Stop	p	t		ṭ		k	ʔ
	b	d		ḍ		g	
Affricate					c		
					j		
Sibilant				s			
Nasal	m	n		ṇ		ṅ	
Lateral		l					
Flap		r	ṛ				
Fricative	v			y		h	

Questions

1. Is this transliteration representing the phonemic system of Kuvi without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

17Pengo(BurrowandBhattacharya1970:1)

Vowels

i	ī			u	ū
e	ē			o	ō
		a	ā		

Consonants

p	b	t	d	ṭ	ḍ	c	j	k	g	
		s	z							h
	m		n		ṇ				ŋ	
					ṛ					
					r					
					l					
	v						y			

Questions

1. Is this transliteration representing the phonemic system of Pengo without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

CentralDravidian

18Kolami(Emeneau1961:§1)

Vowels

i	ī			u	ū
e	ē			o	ō
		a	ā		

Consonants

	Labial	Labio-dental	Dental	Post-dental	Retroflex	Palatal	Velar
Stop	p b		t D		ṭ ḍ		k g
Affricate						c j	
Sibilant				s z			
Trill				r			
Lateral				l			
Nasal	m		N				ŋ
Fricative		v				y	

Questions

1. Is this transliteration representing the phonemic system of Kolami without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

19Naikṛi (Thomasiah1986:§1)

Vowels

i	ī			u	ū
e	ē			o	ō
		a	ā		

Consonants

	Labial	Dental	Alveolar	Retroflex	Palatal	Velar	Glottal
Stop	p b	t d		ṭ ḍ		k g	
	ph bh	th dh		ṭh ḍh		kh gh	
Affricate			c		č j		
					jh		
Nasal	m		n			ŋ	
Fricative	v		s				h
Lateral			l	ḷ			
Trill			r				
Semivowel					y		

Questions

1. Is this transliteration representing the phonemic system of Naikri without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

20Parji(BurrowandBhattacharya1953:§1)

Vowels

i	ī			u	ū
e	ē			o	ō
		a	ā		

Consonants

p	b	t	d	ṭ	ḍ	c	j	k	g	
	m		n				ñ		ŋ	
		[s								h]
			r		ṛ					
			l							
	v						y			

Questions

1. Is this transliteration representing the phonemic system of Parji without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

21Ollari(Bhattacharya1957:partI)

Vowels[14]

i	ī			u	ū
e	ē			o	ō
		a	ā		

Consonants

	Labial	Labio-dental	Dental	Post-dental	Retroflex	Palatal	Velar
Stop	p b		t d		ṭ ḍ		k g
Affricate				ts dz		c j	
Nasal	m		n			(ñ)	ŋ
Rolled				r			
Flapped				ṛ			
Lateral				l			
Fricative		v				y	
Sibilant				s z			

Questions

1. Is this transliteration representing the phonemic system of Ollari without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

22Gadaba(Bhaskararao1998:329ff.)

Vowels

i	ī			u	ū
e	ē			o	ō
		a	ā		

Consonants

	Labial	Dental	Retroflex	Palatal	Velar
Stops:					
Voiceless	p	t	ṭ	c	k
Voiced	b	d	ḍ	j	g
Nasal	m	n	ṇ		ŋ
Fricative		s			
Trill		r			
Lateral		l			
Glide	v			y	

Questions

1. Is this transliteration representing the phonemic system of Gadaba without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

NorthDravidian

23Kurux(Grignard1924a:1–15.Grignard'sclassificationofconsonantsis givenasitis.)

Vowels

i	ī			u	ū
e	ē			o	ō
		a	ā		

Nasalizedvowels[15]

ĩ	ī̃			ũ	u
ẽ	e			(õ)	o
		(ã)	a		

Consonants

Gutturals	k	kh, k̲h̲	g	gh
Palatals	c	ch	j,y	jh
Cerebrals	ṭ	ṭh	ḍ,ṛ	ḍh,ṛh
Dentals	t	th	d	dh
Labials	p	ph	b	bh
Liquids	l	m	n	r
Sibilants,etc.	s	h	w	

Questions

1. Is this transliteration representing the phonemic system of Kuṛux without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5 Which transliterations of specific phonemes to be revised? (Please suggest the alternate system).

24Malto(Mahapatra1979:19–20)

Vowels

i	ī			u	ū
e	ē			o	ō
		a	ā		

Consonants[17]

	Labial	Dental	Alveolar	Retroflex	Palatal	Velar	Uvular	Glottal
Stop								
Voiceless	p	t		ṭ	c	k	q	
Voiced	b	d		ḍ	j	g		
Nasal	m	n			ñ	ṅ		
Fricative		ð	s				γ	h
Trill			r					
Lateral			l					
Flap				ṛ				
Semivowel	w				y			

Questions

1. Is this transliteration representing the phonemic system of Malto without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

25Brahui[18](Emeneau1962d;Elfenbein1997:798–800,1998:392)

Vowels

i	ī			u	ū
	ē				ō
		a	ā		

Consonants[19]

Stops	p	b	t	d			ṭ	ḍ	k	g	?
Affricate					č	j					
Fricative	f								x	ɣ	h
Spirant			s	z	š	ž					
Nasal	m			n			ṇ			(ŋ)	
Lateral			ł	l							
Flap			r				ṛ				
Semivowel	w		y								

Questions

1. Is this transliteration representing the phonemic system of Brahui without any lose of information?

2. Is this transliteration regular and simple? (For instance, under dot [ṇ, ṭ] is regularly followed to represents retroflex sounds)

3. Is this transliteration technologically enabled? (For instance, it should type and rendered with the available Unicode font without many complexities)

4. Is this transliteration standard according to the present linguistic standard? (Standards like Bureau of Indian Standards)

5. Which transliterations of specific phonemes to be revised? (Please suggest the alternate system)

Grammatical Terminology and Abbreviations

Grammatical terminology	**ABBREVIATIONS SUGGESTED**	**ALTRANTE ABBRIVARION**
Abilitative	ABT	
Ability	ABIL	
Ability Component	AC	
Ablative case	ABL	
Absolute case	ABS	
Accusative	ACC	
Active voice	ACV	
Actual	A	
Addressing term	ADD	
Addressive term	ADDT	
Adjectival participle	ADJP	
Adjectival phrase	ADJP	
Adjectival suffix	ADJL	
Adjective	ADJ	
Adverb	ADV	
Adverb of Manner	ADVM	
Adverbial Noun	ADVN	
Adverbial participle	ADVPR	
Adverbial Phrase	ADVP	
Adverbial suffix	ADVL	
Adverbial suffix	ADVS	
Affirmative	AFF	
Affricate	AFR	
Agent	AG	
Agent –like argument of canonical transitive verb	A	
Agentive	AGT	
Agreement	AGR	
Agreement with object	AGR^{OBJ}	
Agreement with subject	AGR^{SUB}	
Allative	ALL	
Anaphora	ANP	
Anaphoric Deictic category	ADC	
Animate	ANIM	

Anticipative	ANTIP	
Aorist	AOR	
Approximant	APR	
Aspirated	ASP	
Aspirated	ASP	
Aspirated	ASP	
Assertion marker clitic	AMC	
Attributive	ATT	
Attributive Phrase	ATTP	
Augment	AUG	
Auxiliary	AUX	
Auxiliary Verb	AUX. V	
Back rounded vowel	BRV	
Back unrounded vowel	BUV	
Back vowel	BV	
Benefactive	BEN	
Cardinal numeral	CARD	
Case	CA	
Case Marker	CM	
Causal Phrase	CP	
Causative	CAUS	
Causative agent	CUA	
Causative suffix	CAUS	
Century	CENT	
Classifier	CL	
Cleft predicate	CLP	
Clitic	CL	
Clitics	CLT	
Collective	COLL	
Comitative	COMIT	
Comitative	COM	
Comparative	CAMP	
Comparative	COMPAR	
Comparative	COMP	
Complement	COMP	
Complement	COMPL	
Complement Noun Phrase	COMPLNP	
Complementizer	COMP	
Completive	COMPLET	
Complex Verb	COMPLV	
Compound verb	COMPV	

Concessive	CONC	
Concomitative	CONCON	
Conditional	COND	
Conditional participle	CNDP	
Conjunction	CONJ	
Conjunction Coordinator	CCD	
Conjunction Sub ordinator	CC.CCS	
Conjunctive participle	CNP	
Conjunctive participle	CP	
Conjunctive participle marker	CPM	
Connective	CON	
Connective	CON	
Consonant	C	
Consonant Vowel Consonant Consonant	CVCC	
Continuous aspect	CONT	
Coordinator	COORD	
Copula	COUP	
Copular	COP	
Correlative	CORR	
Dative	DAT	
Dative case	DAT	
Dative subject construction	DSC	
Debitive	DEB	
Declarative	DECL	
Defective verb	DEFV	
Definite marker	DEF	
Deictic marker	DM	
Demonstrative	DEMO	
Demonstrative	DEM	
Demonstrative base	DB	
Demonstrative Pronoun	DPN	
Dental	DEN	
Derivative Suffix	DS	
Derived Noun	DN	
Determiner	DET	
Dialectal	DIAL	
Direct object	DO	
Directional	DIR	
Disjunction	DIS	
Distributive	DIST	

Dubitative	DUB	
Durative	DUR	
Emotive predicate morpheme	EPM	
Emphasis	EMPH	
Emphatic	EMPH	
Emphatic marker	EMPH	
Emphatic particle	EMP	
Emphatic plural	EMF	
Epenthetic	EPEN	
Epicene plural	EPL	
Ergative	ERG	
Evidentiality	EV	
Example	EG	
Exclusive (first person plural pronoun)	EXCL	
Expletive	EXP	
Extended Predicate	EP	
External Sandhi Rules	ER	
Feminine	F	
Feminine	FEM	
Final complementizer	FC	
Finite	FIN	
First person	1	
First Person Singular	1PRE.SG	
Focus	FOC	
Fricatives	FRI	
Future	FUT	
Future participle	FUTP	
Future tense	FUT	
Gender	GEND	
Gender and Number	G-N	
Gender Marker	GM	
Generic possession marker	GPM	
Genitive	GEN	
Genitive case	GEN	
Glide	GL	
Grammaticalized	GRTD	
Habitual	HAB	
High vowel	HV	
Honorific	HON	
Hortative	HORT	

Human	HUM	
Imperative	IMP	
Imperative	IMP	
Imperative mode	IMPM	
Imperfective	IPFV	
Imperfective aspect	IMPERF$_1$	
Imperfective aspect	IMPERF$_2$	
Impersonal	IMPERS	
Impersonal Negative Nominal	INN	
Impersonal Suffix	IMPL	
Inanimate	INANIM	
Inclusive (first person plural)	INCL	
Incremental vowel	INCRV	
Indefinite	INDF	
Indefinite	INDEF	
Indefinite Plural	INDEFPL	
Indicative	IND	
Indicative	INDI	
Indirect object	ID	
Indirect Object	IO	
Inferential	INFER	
Infinitive	INF	
Infinitive (-**ān**)	INFIN$_2$	
Infinitive (-uka)	INFIN$_1$	
Inflectional phrase	IP	
Initial complementizer	IC	
Injunctive	INJ	
Instrumental case	INST	
Intensifier	INTENS	
Intentive	INTT	
Inter junction	INTJ	
Interjection	INT	
Interrogative	INTERRROG	
Interrogative particle	IP	
Interrogative pronoun	INP	
Intransitive	INTR	
Intransitive	INTR	
Intransitive	IT	
Involitive verb form	INVOL	
Irregular	IRREG	

Known	KN	
Labial	LAB	
Lateral	LAT	
Limitative	LIM	
Literally	LIT	
Locative case	LOC	
Long vowel Consonant	VC	
Low	LO	
Marker	MKR	
Marker	MAR	
Masculine	M	
Masculine (gender)	MAS	
Meaning	MEAN	
Mediative	MED	
Middle	MID	
Mid-honorific	MID-HON	
Modal	MOD	
Modifier	MOD	
Nasal	N	
Necessity	NECES	
Negative	NEG	
Negative Particle	NEGPART	
Neuter	NEU	
Neuter	NEUT	
Neuter Singular	NEUT.SG	
Nominalising suffix	NOML	
Nominalizer	NOZR	
Nominalizer /nominalization	NMLZ	
Nominative case	NOM	
Non future	NONFUT	
Non honorific	NONHON	
Non Honorific	NON.HON	
Non human	NH	
Non masculine	NM	
Non- Masculine	NON-MASC	
Non- Masculine	NONMAS	
Non-honorific	NON-HON	
Nonpast	NP	
Non-specific	NS	
Noun	N	
Noun Lock	NST	

Noun Phrase	NP	
Noun Proper	NNP	
Number	NO	
Number	NO	
Numeral	NUM	
Object	O	
Object	OBJ	
Object agreement marker	OAM	
Objective case	OBC	
Obligation	OBLIG	
Obligative	OBLIG	
Oblique	OBL	
Oblique form	OBL	
Oblique object	OO	
Onomatopoetic	ONO	
Optative	OPT	
Ordinal numeral	ORD	
Overt versus zero case	OZC	
Part	PT	
Participial relative clause	PRC	
Participle	PTC	
Participle	PARTI	
Participle	PTCPL	
Particle	PART	
Particle Default	PRP	
Passive	PASS	
Past	PST	
Past participle	PP	
Past Perfective participle	PPP	
Past tense	PST	
Path case	PATH	
Perfect	PRF	
Perfect	PERF	
Perfective	PERF	
Perfective	PFT	
Perfective aspect	PERFV	
Perfective aspect (-irikk-)	PERF$_1$	
Perfective aspect (-iṭṭuṇṭə)	PERF$_2$	
Perfective Participle	PPL	
Performative Component	PC	

Periphrastic	PERIPH	
Permission	PERMIS	
Permissive	PMS	
Person	P	
Person	PER	
Person Marker	PM	
Personal	PERS	
Personal Affirmative finite Construction	PAF	
Personal ending	PE	
Personal Negative Nominal	PNN	
Phrase	PHR	
Plural	P	
Plural	PL	
Polite	POL	
Possessive	POS	
Possessive	POSS	
Possible/possibility	POSS	
Postposition	PP	
Postpositional phrase	POSTP	
Potential	POT	
Predicate	PRED	
Predicative	PRED	
Preface	PREF	
Present	PRS	
Present	PRES	
Present continues	PC	
Present Future	PRES FUT	
Present perfect continues	PPC	
Present tense	PRES	
Progressive aspect	PROG	
Prohibition	PROHIB	
Prohibitive	PROH	
Pronominal Suffix	PRONS	
Pronominal suffix	PS	
Pronoun	PRON	
Prospective	PROS	
Proximate	PROX	
Proximate	PROX	
Purposive	PURP	
Quantifier	Q	

Quantifier Cardinal	QTC	
Quantifier General	QTF	
Quantifier Ordinal	QTO	
Quantitative adjective	QADJ	
Question	QUES	
Question particle/marker	Q	
Quotative	QUOT	
Quotative Participle	QP	
Reciprocal	RECP	
Reduplication	RED	
Reduplication	REDUP	
Reduplication item	REDU	
Reflexive	REFL	
Reflexive	REF	
Reflexive	REFL	
Relative	REL	
Relative clause	RC	
Relative participle	RP	
Relative participle	RELPTCPL	
Relative participle Suffix	RPS	
Reportative clitic	REPCLT	
Reportative particle	REPORT	
Resultative	RES	
Retroflex	RET	
Retroflex	RET	
Root	RT	
Second person	2	
Section	SEC	
Segmental	SEG	
Self affective	SELFAFF	
Self benefactive	SELFBEN	
Sentence	S	
Sentential Relative clause	SRC	
Simple	SIMP	
Singular	SL	
Singular	SG	
Sociative	SOC	
South Asia/South Asian	SA	
Spoken	SP	
Stative	STAT	
Stem	ST	

Subject	S	
Subject	SUB	
Subject-object-verb	SOV	
Subjunctive	SUBJ	
Subjunctive	SBJV	
Suffix	SUF	
Suffix	SFX	
Suggestive	SUG	
Superlative marker	SUP	
Surprise Verbal form	SURP	
Tense Implied Relative Participle	TIRP	
Tense Marker	TM	
Third person	3	
Third person Feminine Singular	3.PER.FEM.SG	
Third person Masculine Singular	3.PER.MAS.SG	
Topic	TOP	
Topicalized complement clause marker	TCCM	
Transitive	TR	
Transitive	TRANS	
Ultra-honorific- pronoun	UHON	
Un know	UNKN	
Unaspirated	UNASP	
Variant	VAR	
Verb	VB	
Verb Finite	VF	
Verb Gerund	VG	
Verb intransitive	VINTR	
Verb intransitive	VI	
Verb non Finite	VNF	
Verb Phrase	VP	
Verb Stem	VST	
Verb transitive	VTR	
Verb transitive	VT	
Verb/Vowel	V	
Verbal Base	VB	
Verbal compound	VEP	
Verbal noun	VN	

Verbal participle	VERB PART	
Verbal reciprocal	VREC	
Verbal reflexive	VR	
Vocative	VOC	
Voiced	VOD	
Voiced	VD	
Voiced plosive	B	
Voiceless	VOL	
Voiceless	VL	
Voiceless plosives	P	
Volitive optative	VOLOPT	
Yes/no question	Y/N Q MKR	

Glossing Standards

Linguistic glossing is a technique for linguistic analysis and a standard way of presentation of linguistic data in any level. There can be number of levels in glossing according to the purpose and levels of the analysis. An ideal glossed text in Malayalam is given below with different levels of glossing. First is the linguistic text under analysis is generally called object language. Second is the morphemic glossing of each grammatical element. Third is the subcategory glossing. Forth is the grammatical category glossing. Fifth is the phrasal category glossing. Sixth is the meaning of the object language in English. Except first and sixth rest are called the metalanguages. Metalanguage may be morphological, sub categorical, grammatical categorical or phrasal depending upon the level of analysis. See an ideal glossed text in Malayalam. See part three for the standard abbreviation of metalanguage.

Example 2.1 An ideal glossing

1	**Object language**	kaṟu-	tta	pṭṭi	**ve**ḷu-	tta	pšuv-	ne	kaṭi-	ccu
2	**morphemic glossing**	Black-	ADJL	dog	white-	ADJL	Cow-	ACC	bite-	PST
3	**subcategory glossing**	ADJ		N	ADJ		N	CA	V	
4	**Grammatical category glossing**	S			O				V	
5	**Phrasal category glossing**	NP			VP					
6	**Meaning of the text**	"Black dog beaten the white cow"								

There can be number of ways of glossing according to the level and purpose of the analysis and presentation. The *LeGRu* has proposed ten rules for glossing with additional optional. Ten rules can be discussed with the data from Malayalam and see which is more appropriate to Dravidian languages in general and particularly to Malayalam; they are;

Rule 1: Word- by-word alignment

Word by word glossing is the primary way of glossing. Regarding the primary level word by word glossing LeGRu suggests that, "In word by word alignment interlinear glosses are left- aligned vertically, word by word" See an example of word by word glossing in Malayalam (2.1).

Example 2.1:

rāman	sītaye	sn**ēhikkunnu**
Raman	*Sita*	*loves*

Rama loves Sitha

This is a simple word by word glossing. The glossing has done only at the level of word. Since Dravidian languages are rich in morphology, one word consists number of morphemes by both inflectional and derivation process. The above Malayalam example shows that Rule 1 is not adequate enough to analysis the rich morphology of the agglutinative languages like Dravidian. This glossing cannot represent many of the derived and inflected grammatical categories in Malayalam. Inflection of the noun *sīta* for accusative case suffix **–e** and the inflection of *snēham* "love" for transitive suffix ***-kk-*** and present tense suffix ***-unnu*** cannot be analyzed by this glossing. Therefore we have to apply the Rule No.2 of glossing.

Rule 2: Morpheme –by-morpheme correspondence

Regarding the morpheme by morpheme glossing LeGRu) stated that "Segmentable morphemes are separated by hyphens, both in example and in the gloss. There must be exactly the same number of hyphens in the example and in the gloss". See the Malayalam example of this Rule2 (1.2.2).

Example 1 2.2

avan	nallav-	an	skūḷ-	il	pōk-	um
he	*good-*	*M*	*school-*	*LOC*	*go-*	*FUT*

" He is good and will go to school"

Compares to the *Rule 1* this glossing is adequate enough to analysis the inflectional and derivational forms of Malayalam. But there are two forms in this object language which are not properly explained. The form for good is *nalla,* a masculine suffix - *an* is suffixed on it. When *nalla* is inflected for *–an* it realize as *nalla(v)an* by morphophonemic change. An additional consonant *-v-* is inserted between the stem and the suffix. But this rule cannot explain the sound change in morphemic boundary. This is one of limitation of this rule 2. Another problem is that, the form *-an* which is glossed as agreement suffix have some other properties of singular and gender is called the values of agreement. But this rule is not adequate enough to represent such factors. Therefore we can go for the Rule.3 for further solution.

Rule 3: Grammatical category labels

LeGRu states that "Grammatical morphemes are generally rendered by abbreviates grammatical category labels, printed in upper case letters (usually small capitals).

Example 1.3.1

avan-	um	avaḷ	um-	skūḷ	**lēkk ə**	p**ō-**	i
2SL-	*CONJ*	*2SL*	*CONJ*	*school*	*LOC*	*V-*	*PST*
he-	*with*	*she-*	*with*	*school*	*towards*	*go-*	*PST*

He and she went to school

This glossing is giving three kind of information; morpheme by morpheme segments, labels of grammatical category and possible meaning of each category. The form *avan* "he" and the form *avaḷ* "she" means not only indicates feminine and masculine only, but represent the number and person also. But the technique of this rule is not adequate enough to represent such additional information. Therefore we have to look the Rule 4 for further solution.

Rule 4: One to many correspondences

LeGRu states that “when a single object –language element is rendered by several meta language elements (words or abbreviations), these are separated by periods”. See the example from Malayalam;

Example 2.4.1

avan-	um	avaḷ	um-	skūḷ	il**ēkkə**	p**ō-**	i
2SL.M-	*CONJ*	*2SL.F*	*CONJ*	*school*	*LOC.DIR*	*v-*	*pst*
he-	*with*	*she-*	*with*	*school*	*towards*	*go-*	*pst*

He and she went to school

Here what is missed in the above glossing, i.e. the element of masculine in *avan-* “second person singular masculine (2SL.M), avaḷ “second person feminine singular” also glossed (2SL.F) and the direction indication of *ilēkkə* “towards” also can be glossed.

Rule Number 4A, 4B and 4C is not applicable to Dravidian language, but the Rule 4D can treat the morphophonological process in Dravidian languages.

Rule4 C. Non- segmentation

If an author is not intended to segment some elements LeGRu (:5) stated that “if an object-language element is formally and semantically segmentable, but the author does not want to show the formal segmentation (because it is irreverent and /or to keep the text intact), the colon may be used”. This is most relevant in the grammaticalized items in Dravidian languages. If the analysis is not intended to provide evolutionary information it is much relevant in Modern Malayalam;

Example 2.4C.1

avaḷ	kaṭa-	il	p**ō-**	i
DEM:F	shop-	*LOC*	*go-*	*PST*

She went to shop

Here the above example avaḷ "she" is an grammaticalized item by the combination of demonstrative form a and aḷ "feminine".

Rule 4D. Morphophonology

LeGRu stated that "if a grammatical property in the object-language is signaled by a morphophonological change (ablaut, mutation, tone alternation, etc.), the backlash is used to separate the category label and the rest of the gloss".See the below Malayalam example;

Example 2.4D.1

rāmu-	vir̠e	
Ramu\	*GEN*	*Genitive:* r̠e

Ramus

pū	vum	
flower\	*CONJ*	*Conjunct: um*

Even though this glossing is showing the morophophonological process, it is not giving the information of the source of morphophonological change. Therefore the source of morphophonemic change also indicated.

Rule 5.Persion and number labels

LeGRu (:6) stated that "Person and number are not separated by a period when they occur in this order";

Example 2.5.1

avan	nalla-	van
he	*good-*	*M.SL*

He is good

Example2.5.2

avaḷ	nalla-	vaḷ
she	*good-*	*F.SL*

He is good

Rule 6: Non-overt elements

Regarding the non overt elements LeGRu suggests that “If the morpheme-by morpheme gloss contains an element that does not correspond to an overt element in the example, it can be enclosed in square brackets” .See the examples from Malayalam in 2.6.1 and 2.6.2.

Example 2.6.1

avan	pšuv-	ne	aṭi-	ccu
he [NOM.SG]	cow-	*DAT*	*beat-*	*PST*

He beaten the cow

OR

Example 2.6.2.

avan-ø	pšuv-	ne	aṭi-	ccu

he -NOM.SG	cow-	*DAT*	*beat-*	*PST*

He beaten the cow

Rule 7: Inherent categories

Inherent, no overt categories such as gender may be indicated in the gloss, but a special boundary symbol, the round parenthesis, is used.

atə	oru	pšu	āṇə
that (NEU.SG)	one	cow	*be-* PRES

"That is a cow"

Rule 8: Bipartite elements: [Not applicable in Dravidian Languages]

Rule 9: Infixes: [Not applicable in Dravidian Languages]

Rule 10: Reduplication

Reduplication is a frequent feature in Dravidian languages. LeGRu suggests that "Reduplication is treated similarly to affixation, but with a tilde (instead of an ordinary hyphen) connecting the copied element to the stem.

valiya~	valiya	kārya-	ňňaḷ
big~	RED	subject-	PL

“Big big things”

kiḷi-	kaḷ	kala~	pila	kara-	ññu
bird-	PL	? ~	RED	cry-	*PST*

“Birds are crying like….”

2.1. Technology

Microsoft word can be exclusively used for the above discussed glossing techniques. In addition to the above discussed transliteration (chapter 1) use of table makes the glossing much easier and systematic. Table can be inserted according to the level glossing given below:

	Object language					
	Metalanguage					
	Meaning					

In the above table first row can be used for object language, second is for metalanguage and third can be used for meaning. After glossing lines of the table can be erased by changing the properties of the table;

Supporting materials

Lehmann, Christian. 2004a. Interlinear morphemic glossing. http://www.unierfurt.de/sprachwissenschaft/personal/lehmann/CL_Publ/IMG.PDF.

Lehmann, Christian. 2004b. Interlinear morphemic glossing. In Geert Booij, Christian Lehmann, Joachim Mugdan & Stavros Skopeteas (eds.), Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. 2. Halbband, 1834-57. Berlin: Mouton de Gruyter.

Lieb, Hans-Heinrich & Sebastian Drude. 2000. Advanced glossing: A language documentation format. http://www.mpi.nl/DOBES/documents/Advanced-Glossing1.pdf.

Schultze-Berndt, Eva. 2006. Linguistic annotation. In Jost Gippert, Nikolaus Himmelman & Ulrike Mosel (eds.), *Essentials of language documentation*, 213-52. Berlin: Mouton de Gruyter.

Bird Steven & Gary Simons. 2003, Seven Dimensions of Portability for Language Documentation and Description. Language 79: 557-582.

Bickel, B., Comrie, B. and Haspelmath, M., 2008. The Leipzig Glossing Rules. Conventions for interlinear morpheme by morpheme glosses. *Revised version of February*.

Annamalai, E. and Steever, S. B. 1998. Modern Tamil. In Steever, S. B. (ed.), The Dravidian Languages, 100-128. London: Routledge.

R. E. Asher, T. C. Kumari. Routledge, 1997 - Foreign Language Study - 491 pages

Acharya Balkrishna, 2008. .... Flora of Coorgu (Kodagu). Karnataka, India. Vimsat .... Saldanha, C. J. & Nicholson, 1976. The Flora of Hassan

Kapp, Dieter B. 1987. "Centralized Vowels in Alu Kurumba." In Journal of the American Oriental Society, 107 , no. 3: 409--426. American Oriental Society.

The Iruḷa language, Volumes 1-3. Front Cover. KamilZvelebil. Harrassowitz, 1973 - Irula language - 64 pages.

Emeneau, M.B. 1944. Kota Texts California: University of California Press

Emeneau, M. B. (1958), "Oral Poets of South India: Todas", Journal of American Folklore, 71 (281): 312–324

Emeneau, M. B. 1984. "[Untitled]." In Language, 60 , no. 3: 675--676. Linguistic Society of America

Shalev, E., Geslevich, Y. and Ben-Ami, M. (1994) Induction of pre-ovulatory luteinizing hormone surge by gonadotrophin-releasing hormone agonist for women at risk for developing the ovanan hyperstimulation syndrome Hum. Reprod., 9, 417-419

A Grammar of the Kannada Language: Comprising the Three Dialects of the Language (ancient, Medieval and Modern) ..... Kannada / S.R.Sridhar. Sridhar, S. R., 1950- , 1990 ...

"Badaga language not a dialect of Kannada, claims French ... Jump up ^ Paul Hockings, Christiane Pilot-Raichoor (Reprint 1992).

"Tourism in ... "Tulu Nadu: The Land and its People by Dr. Neria H. Hebbar". Boloji. ... D.N.S. Bhat (1998). Sanford B

The Koragalanguage. Poona, 1971. Bhat 1971a - D.N.S. Bhat. Havyaka.

Krishnamurti, Bh. 1998. Telugu. In Steever, Sanford B. (ed.), The Dravidian Languages, 202-240. London and New York: Routledge

Umamaheshwar Rao. 1987. A comparative study of the Gondi dialects (with special reference to phonology and morphology). Osmania University.

Krishnamurti, Bhadriraju. 1969. Konda or Kubi. A Dravidian language. Hyderabad: Government of Andhra Pradesh, Tribal Cultural Research and Training Institute

Winfield, W. W. 1928. A Grammar of the Kui Language. (Bibliotheca Indica, 245.) Calcutta: Asiatic Society of Bengal.

Israel 1979. Text; BibTeX; RIS; MODS. Israel, M. 1979. A Grammar of the Kuvi Language. Trivandrum, India: Dravidian Linguistics Association

Burrow and Bhattacharya 1970. Text; BibTeX; RIS; MODS. Burrow, Thomas and Bhattacharya, S.1970. The Pengo Language. London: Oxford University Press.

M. B. Emeneau, A Dravidian ... Oxford 1961. ... Emeneau,Kolami Emeneau, M. B. , Kolami, a Dravidian Language.

Thomasiah, K. 1986 Naikri dialect of Kolami: Descriptive and comparative study. Annamalai University PhD dissertation. Konda (Dr) Krishnamurti, Bhadriraju

NTRODUCTION Parji, Kolami, and Ollari (also known as Konekor Gadaba) ... PARJI A cursory search of Burrow and Bhattacharya's (1953)

Ollari: A Dravidian Speech. Front Cover. Sudhibhushan Bhattacharya. Manager of Publications, 1957 - Gadaba language (Dravidian)

Bhaskararao, Peri. 1998. Gadaba. In Steever, Sanford B. (ed.), The Dravidian Languages, 328-355. London: ...

Grignard 1924, A Grammar of the Oraon Language and Study in Oraon Idiom, 1924, 317, grammar

Mahapatra, BR 1979. Malto: An Ethnosemantic Study. Mysore: Central Institute of ... A ReferenceGrammar of Colloquial Burmese. London: Oxford University.

Emeneau, Murray B. 1962. Brahui and Dravidian Comparative Grammar. (University of California Publications in Linguistics, 27.) Berkeley: University of California Press.

Leipzig, last change: May 31, 2015
Further updates will be managed by the Committee of Editors of Linguistics Journals.

The Leipzig Glossing Rules:

Conventions for interlinear morpheme-by-morpheme glosses

About the rules

The Leipzig Glossing Rules have been developed jointly by the Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology (Bernard Comrie, Martin Haspelmath) and by the Department of Linguistics of the University of Leipzig (Balthasar Bickel). They consist of ten rules for the "syntax" and "semantics" of interlinear glosses, and an appendix with a proposed "lexicon" of abbreviated category labels. The rules cover a large part of linguists' needs in glossing texts, but most authors will feel the need to add (or modify) certain conventions (especially category labels). Still, it will be useful to have a standard set of conventions that linguists can refer to, and the Leipzig Rules are proposed as such to the community of linguists. The Rules are intended to reflect common usage, and only very few (mostly optional) innovations are proposed.

We intend to update the Leipzig Glossing Rules occasionally, so feedback is highly welcome.

Important references:

Lehmann, Christian. 1982. "Directions for interlinear morphemic translations". *Folia Linguistica* 16: 199-224.

Croft, William. 2003. *Typology and universals.* 2nd ed. Cambridge: Cambridge University Press, pp. xix-xxv.

The rules

(revised version of February 2008)

Preamble

Interlinear morpheme-by-morpheme glosses give information about the meanings and grammatical properties of individual words and parts of words. Linguists by and large conform to certain notational conventions in glossing, and the main purpose of this document is to make the most widely used conventions explicit.

Depending on the author's purposes and the readers' assumed background knowledge, different degrees of detail will be chosen. The current rules therefore allow some flexibility in various respects, and sometimes alternative options are mentioned.

The main purpose that is assumed here is the presentation of an example in a research paper or book. When an entire corpus is tagged, somewhat different considerations may apply (e.g. one may want to add information about larger units such as words or phrases; the rules here only allow for information about morphemes).

It should also be noted that there are often multiple ways of analyzing the morphological patterns of a language. The glossing conventions do not help linguists in deciding between them, but merely provide standard ways of abbreviating possible descriptions. Moreover, glossing is rarely a complete morphological description, and it should be kept in mind that its purpose is not to state an analysis, but to give some further possibly relevant information on the structure of a text or an example, beyond the idiomatic translation.

A remark on the treatment of glosses in data cited from other sources: Glosses are part of the analysis, not part of the data. When citing an example from a published source, the gloss may be changed by the author if they prefer different terminology, a different style or a different analysis.

Rule 1: Word-by-word alignment

Interlinear glosses are left-aligned vertically, word by word, with the example. E.g.

(1) Indonesian (Sneddon 1996:237)
Mereka di Jakarta sekarang.
they in Jakarta now
'They are in Jakarta now.'

Rule 2: Morpheme-by-morpheme correspondence

Segmentable morphemes are separated by hyphens, both in the example and in the gloss. There must be exactly the same number of hyphens in the example and in the gloss. E.g.

(2) Lezgian (Haspelmath 1993:207)
Gila abur-u-n ferma hamišaluğ güğüna amuq'-da-č.
now they-OBL-GEN farm forever behind stay-FUT-NEG
'Now their farm will not stay behind forever.'

Since hyphens and vertical alignment make the text look unusual, authors may want to add another line at the beginning, containing the unmodified text, or resort to the option described in Rule 4 (and especially 4C).

Clitic boundaries are marked by an equals sign, both in the object language and in the gloss.

(3) West Greenlandic (Fortescue 1984:127)
palasi=lu niuirtur=lu
priest=and shopkeeper=and
'both the priest and the shopkeeper'

Epenthetic segments occurring at a morpheme boundary should be assigned to either the preceding or the following morpheme. Which morpheme is to be chosen may be determined by various principles that are not easy to generalize over, so no rule will be provided for this.

Rule 2A. (Optional)
If morphologically bound elements constitute distinct prosodic or phonological words, a hyphen and a single space may be used together in the object language (but not in the gloss).

(4) Hakha Lai
a-nii -láay
3SG-laugh-FUT
's/he will laugh'

Rule 3: Grammatical category labels

Grammatical morphemes are generally rendered by abbreviated grammatical category labels, printed in upper case letters (usually small capitals). A list of standard abbreviations (which are widely known among linguists) is given at the end of this document.

Deviations from these standard abbreviations may of course be necessary in particular cases, e.g. if a category is highly frequent in a language, so that a shorter abbreviation is more convenient, e.g. CPL (instead of COMPL) for "completive", PF (instead of PRF) for "perfect", etc. If a category is very rare, it may be simplest not to abbreviate its label at all.

In many cases, either a category label or a word from the metalanguage is acceptable. Thus, both of the two glosses of (5) may be chosen, depending on the purpose of the gloss.

(5) Russian

My	*s*	*Marko*	*poexa-l-i*	*avtobus-om*	*v*	*Peredelkino.*
1PL	COM	Marko	go-PST-PL	bus-INS	ALL	Peredelkino
we	with	Marko	go-PST-PL	bus-by	to	Peredelkino

'Marko and I went to Perdelkino by bus.'

Rule 4: One-to-many correspondences

When a single object-language element is rendered by several metalanguage elements (words or abbreviations), these are separated by periods. E.g.

(6) Turkish
çık-mak
come.out-INF
'to come out'

(7) Latin
insul-arum
island-GEN.PL
'of the islands'

(8) French
aux *chevaux*
to.ART.PL horse.PL
'to the horses'

(9) German
unser-n *Väter-n*
our-DAT.PL father.PL-DAT.PL
'to our fathers'

(10) Hittite (Lehmann 1982:211)
n=an *apedani* *mehuni* *essandu.*
CONN=him that.DAT.SG time.DAT.SG eat.they.shall
'They shall celebrate him on that date.' (CONN = connective)

(11) Jaminjung (Schultze-Berndt 2000:92)
nanggayan *guny-bi-yarluga?*
who 2DU.A.3SG.P-FUT-poke
'Who do you two want to spear?'

The ordering of the two metalanguage elements may be determined by various principles that are not easy to generalize over, so no rule will be provided for this.

There are various reasons for a one-to-many correspondence between object-language elements and gloss elements. These are conflated by the uniform use of the period. If one wants to distinguish between them, one may follow Rules 4A-E.

Rule 4A. (Optional)
If an object-language element is neither formally nor semantically segmentable and only the metalanguage happens to lack a single-word equivalent, the underscore may be used instead of the period.

(12) Turkish (cf. 6)
çık-mak
come_out-INF
'to come out'

Rule 4B. (Optional)
If an object-language element is formally unsegmentable but has two or more clearly distinguishable meanings or grammatical properties, the semi-colon may be used. E.g.

(13) Latin (cf. 7)
insul-arum
island-GEN;PL
'of the islands'

(14) French

aux	*chevaux*
to;ART;PL	horse;PL

'to the horses'

Rule 4C. (Optional)
If an object-language element is formally and semantically segmentable, but the author does not want to show the formal segmentation (because it is irrelevant and/or to keep the text intact), the colon may be used. E.g.

(15) Hittite (Lehmann 1982:211) (cf. 10)

n=an	*apedani*	*mehuni*	*essandu.*
CONN=him	that:DAT;SG	time:DAT;SG	eat:they:shall

'They shall celebrate him on that date.'

Rule 4D. (Optional)
If a grammatical property in the object-language is signaled by a morphophonological change (ablaut, mutation, tone alternation, etc.), the backslash is used to separate the category label and the rest of the gloss.

(16) German (cf. 9)

unser-n	*Väter-n*
our-DAT.PL	father\PL-DAT.PL

'to our fathers' (cf. singular *Vater*)

(17) Irish

bhris-is
PST\break-2SG
'you broke' (cf. nonpast *bris-*)

(18) Kinyarwanda

mú-kòrà
SBJV\1PL-work
'that we work' (cf. indicative *mù-kòrà*)

Rule 4E. (Optional)
If a language has person-number affixes that express the agent-like and the patient-like argument of a transitive verb simultaneously, the symbol ">" may be used in the gloss to indicate that the first is the agent-like argument and the second is the patient-like argument.

(19) Jaminjung (Schultze-Berndt 2000:92) (cf. 11)

nanggayan	*guny-bi-yarluga?*
who	2DU>3SG-FUT-poke

'Who do you two want to spear?'

Rule 5: Person and number labels

Person and number are not separated by a period when they cooccur in this order. E.g.

(20) Italian
and-iamo
go-PRS.1PL (not: *go*-PRS.1.PL)
'we go'

Rule 5A. (Optional)

Number and gender markers are very frequent in some languages, especially when combined with person. Several authors therefore use non-capitalized shortened abbreviations without a period. If this option is adopted, then the second gloss is used in (21).

(21) Belhare

ne-e	*a-khim-chi*	*n-yuNNa*
DEM-LOC	1SG.POSS-house-PL	3NSG-be.NPST
DEM-LOC	1sPOSS-house-PL	3ns-be.NPST

'Here are my houses.'

Rule 6: Non-overt elements

If the morpheme-by-morpheme gloss contains an element that does not correspond to an overt element in the example, it can be enclosed in square brackets. An obvious alternative is to include an overt "Ø" in the object-language text, which is separated by a hyphen like an overt element.

(22) Latin

puer	or:	*puer-Ø*
boy[NOM.SG]		boy-NOM.SG
'boy'		'boy'

Rule 7: Inherent categories

Inherent, non-overt categories such as gender may be indicated in the gloss, but a special boundary symbol, the round parenthesis, is used. E.g.

(23) Hunzib (van den Berg 1995:46)

oz#-di-g	*xõxe*	*m-uq'e-r*
boy-OBL-AD	tree(G4)	G4-bend-PRET

'Because of the boy the tree bent.'

(G4 = 4th gender, AD = adessive, PRET = preterite)

Rule 8: Bipartite elements

Grammatical or lexical elements that consist of two parts which are treated as distinct morphological entities (e.g. bipartite stems such as Lakhota *na-xʔų* 'hear') may be treated in two different ways:

(i) The gloss may simply be repeated:

(24) Lakhota
na-wíčha-wa-xʔų
hear-3PL.UND-1SG.ACT-hear
'I hear them' (UND = undergoer, ACT = actor)

(ii) One of the two parts may be represented by a special label such as STEM:

(25) Lakhota
na-wíčha-wa-xʔų
hear-3PL.UND-1SG.ACT- STEM
'I hear them'

Circumfixes are "bipartite affixes" and can be treated in the same way, e.g.

(26) German
ge-seh-en or: *ge-seh-en*
PTCP-see-PTCP PTCP-see-CIRC
'seen' 'seen'

Rule 9: Infixes

Infixes are enclosed by angle brackets, and so is the object-language counterpart in the gloss.

(27) Tagalog
b<um>ili (stem: *bili)*
<ACTFOC>buy
'buy'

(28) Latin
reli<n>qu-ere (stem: *reliqu-)*
leave<PRS>-INF
'to leave'

Infixes are generally easily identifiable as left-peripheral (as in 27) or as right-peripheral (as in 28), and this determines the position of the gloss corresponding to the infix with respect to the gloss of the stem. If the infix is not clearly peripheral, some other basis for linearizing the gloss has to be found.

Rule 10: Reduplication

Reduplication is treated similarly to affixation, but with a tilde (instead of an ordinary hyphen) connecting the copied element to the stem.

(29) Hebrew
yerak~rak-im
green~ATT-M.PL
'greenish ones' (ATT = attenuative)

(30) Tagalog
bi~bili
IPFV~buy
'is buying'

(31) Tagalog
b<um>i~bili
<ACTFOC>IPFV~buy
'is buying' (ACTFOC = Actor focus)

Appendix: List of Standard Abbreviations

1	first person
2	second person
3	third person
A	agent-like argument of canonical transitive verb
ABL	ablative
ABS	absolutive
ACC	accusative
ADJ	adjective
ADV	adverb(ial)
AGR	agreement
ALL	allative
ANTIP	antipassive
APPL	applicative
ART	article
AUX	auxiliary
BEN	benefactive
CAUS	causative
CLF	classifier
COM	comitative
COMP	complementizer
COMPL	completive
COND	conditional
COP	copula
CVB	converb
DAT	dative
DECL	declarative
DEF	definite

DEM	demonstrative
DET	determiner
DIST	distal
DISTR	distributive
DU	dual
DUR	durative
ERG	ergative
EXCL	exclusive
F	feminine
FOC	focus
FUT	future
GEN	genitive
IMP	imperative
INCL	inclusive
IND	indicative
INDF	indefinite
INF	infinitive
INS	instrumental
INTR	intransitive
IPFV	imperfective
IRR	irrealis
LOC	locative
M	masculine
N	neuter
N-	non- (e.g. NSG nonsingular, NPST nonpast)
NEG	negation, negative
NMLZ	nominalizer/nominalization
NOM	nominative
OBJ	object
OBL	oblique
P	patient-like argument of canonical transitive verb
PASS	passive
PFV	perfective
PL	plural
POSS	possessive
PRED	predicative
PRF	perfect
PRS	present
PROG	progressive
PROH	prohibitive
PROX	proximal/proximate
PST	past
PTCP	participle
PURP	purposive
Q	question particle/marker
QUOT	quotative
RECP	reciprocal
REFL	reflexive

REL	relative
RES	resultative
S	single argument of canonical intransitive verb
SBJ	subject
SBJV	subjunctive
SG	singular
TOP	topic
TR	transitive
VOC	vocative

References

Fortescue, Michael. 1984. *West Greenlandic.* (Croom Helm descriptive grammars) London: Croom Helm.

Haspelmath, Martin. 1993. *A grammar of Lezgian.* (Mouton Grammar Library, 9). Berlin - New York: Mouton de Gruyter.

Lehmann, Christian. 1982. "Directions for interlinear morphemic translations". *Folia Linguistica* 16: 199-224.

Schultze-Berndt, Eva. 2000. *Simple and complex verbs in Jaminjung: A study of event categorization in an Australian language.* Katholieke Universiteit Nijmegen Ph.D. Dissertation.

Sneddon, James Neil. 1996. *Indonesian*: *A comprehensive grammar.* London: Routledge.

van den Berg, Helma. 1995. *A Grammar of Hunzib.* (Lincom Studies in Caucasian Linguistics, 1.) München: Lincom Europa.

The linguistic example*

David J. Weber
Summer Institute of Linguistics

Good language descriptions liberally illustrate their claims with examples. The author must select and order examples, and provide accompanying information. The example may include a reference number, the example in multiple forms (phonetic, phonemic, morphemic or morphophonemic, written), brackets and categories, glosses, translation, punctuation, functional annotations, grammatical judgements, subscripts, empty categories, ellipses marking, information about the author and language variety, attention-directing mechanisms, and so forth. Formatting these diverse sorts of information is a non-trivial task; suggestions are given for "best practice." The delivery of documents on screens (rather than on paper) makes possible some dynamic enhancements such as inspecting an example's textual context, toggling on/off various types of information, controlling highlighting and conflation.

1. Introduction

The linguistic literature is populated by a menagerie of "specials": tables, trees, maps, HPSG's rectilinear attribute-value matrices, RG's curvaceous stratal diagrams, and so forth. Among these, the most important for baseline language descriptions is the EXAMPLE: words, phrases, sentences, and text fragments used to illustrate claims made about the language under consideration.

Good language descriptions liberally illustrate their claims with examples, ideally ones drawn from natural discourses of diverse genre.[1] Good examples, well deployed, are a major factor in making a grammar good.

Think of a grammar like a Museum of Fine Art. The collection is laid out topically in galleries, so we might come to the Gallery of Relative Clauses. Each piece is tastefully framed, and light is provided to bring the best out of each piece. One can stop and ponder, but is eventually drawn from one awe-inspiring piece to another.

Studies in Language 30:2 (2005), 445–460.
ISSN 0378–4177 / E-ISSN 1569–9978 © John Benjamins Publishing Company

A museum, of course, is not a warehouse, which may contain an incredible store of pieces, each carefully shelved according to some organizational scheme. A grammar must be more — much more — than simply an annotated data catalog. But keep in mind that some day the examples may be appreciated more than the author's fine words giving some clever analysis or theory. With time the claims may become uninteresting, as concerns and perspectives change, but examples remain as near-primary evidence. (Of course, for a long time to come readers will appreciate surrounding text that helps them understand the example, such as a description of the context in which the example was uttered.)

Perhaps the most important reason to pay careful attention to examples is that they stimulate and exploit abduction. Abduction is inference to the best explanation: for an array of phenomena one abduces (guesses) possible explanations. Each explanation is "probably true" in proportion to how well it accounts for the array of phenomena, with higher probability for explanations that cover diverse phenomena.

An example (a phenomenon) normally follows the claim (explanation) that it illustrates. Nonetheless, it can stimulate abduction; indeed, the examples should be ones from which a reader might have abduced the claim. Further examples can strengthen the claim, especially if these are diverse, either within the example or used in different ways in different contexts. Diverse examples strengthen the claim because diversity makes it more challenging to abduce a better competing explanation.

2. Selection, order, framing, enrichment

When writing a grammar, the author must select, order, frame, and enrich examples, as discussed in the following sections.

2.1 Selection

On what basis should the author select examples for incorporation into a text? Here are some suggestions:

- First and foremost, the examples must illustrate the claim being made.
- The examples must be sound, preferably ones spoken or written by a native speaker in natural discourse. Such examples should always be preferred to elicited examples. Avoid examples that presume a contrived context, if possible.

- Choose a set of examples that illustrate a range of uses. Otherwise the reader might think that the claim holds only for a narrow range of uses.
- For the same reason, choose a set of structurally diverse examples. An example that is highly similar to another adds virtually nothing.
- Finally, when all else is equal, choose examples that are culturally interesting.

There are other kinds of concern:

1. Beware of various types of bias. To take one case, the examples of some grammars reflect gender bias: subjects (agents) are more likely to be male than female; females are more likely to appeared as objects (undergoers) or in oblique roles; examples in which an undergoer is adversely affected by an agent (like "X struck Y") are more likely to have a male agent and a female undergoer than the other way around; and so forth. Likewise, the examples may reflect racial prejudice, or even hatred for members of another human group.

   Such biases might reflect attitudes held by the speakers of the language under consideration, or at least of those who provided the data. It may be residual in the language's literature (both oral and written). Or it might simply reflect what speakers regard as note-worthy.

   A corpus formed from real, natural instances of language use will reflect the prejudices of the speakers. As examples are taken from the corpus and incorporated into a grammar, the grammar will reflect those prejudices.

   It will do so, that is, unless the author is aware of the problem and exercises good judgement. I am not advocating the imposition of some radical notion of political correctness, but the problem can be reduced by the judicious choice of examples and occasionally by small adjustments (approved by a native speaker).
2. A grammar writer should bear in mind that the examples in a grammar will project an image of the speakers of the language and their culture, one that may be seen around the world (if made available on the web) and by speakers of the language, now and in future generations. So grammar writers should take care not to expose — or inadvertantly perpetuate — prejudices and other aspects of the culture that might embarass its speakers.
3. If you work from a corpus — as you should — watch out for examples that might compromise individuals or groups. Bear in mind that some people love to tell and retell stories that embarass or damage their enemies. These stories may be ethnographically interesting but quite inappropriate for a grammar, for the reasons just mentioned.

4. Examples should be avoided that could limit the usefulness of the grammar. For example, when a Huallaga Quechua grammar describes the use of *na-*, a verb devoid of semantic content, to create suspense, a fine example would be:

   (1) …pullan pagasnaga nasha auquenga.
   '…in the middle of the night the old man *did* him.'

   Following this sentence, for the next 132 words the reader wonders what the old man did to his wife's lover. Then the text ends with:

   (2) Quiquin wañuraycachir auquenga rucsuntash cuchuriycuran.
   Bulsicacurcur apacun.
   'He himself, the old man, killed him and cut off his testicles.
   He put them into his pocket and took them.'

   This nicely illustrates the phenomena but might keep the grammar from being used in, say, a secondary school.

2.2 Order

It is often appropriate to use several examples. When this is so, how should they be ordered? Here are some suggestions:

- Begin with those that best exemplify the claim being supported, those for which the relationship to the claim is most clear.
- Progress from the simplest (usually the shortest) examples to the more complex ones.
- Progress from unmarked cases (vanilla) to more marked ones (rocky road).
- If included at all, put ambiguous cases last. These would be examples for which there is an alternative interpretation that might undercut the example's support for the claim.

2.3 Framing

Think of a cut diamond. Alone, it is valuable and holds a certain charm, but when set in a piece of jewelry, then it becomes truly beautiful.

Examples are like that. Examples must be framed. Generally it does not suffice to make a claim and then simply tack on one or more examples. The reader needs help to see how the example illustrates the claim. For example, consider the following:

> Barasano verbs bear suffixes agreeing with the subject of the clause:
> (3) sĩg-o ĩa-a-bõ ‘The woman sees.’
> one-FEM SEE-PRES-FEM.SG

Now compare this to the following, which provides help explaining how the example relates to the claim:

> Barasano verbs bear suffixes agreeing with the subject of the clause. For example, in (4) the verbal suffix -bõ agrees with sĩg-o ‘woman’:
> (4) sĩg-o ĩa-a-bõ ‘The woman sees.’
> one-FEM SEE-PRES-FEM.SG

In (4) the order is CLAIM-HELP-EXAMPLE. It is sometimes better to put the help after the example. If so, the order would be (1) the CLAIM, (2) some text like “Consider the following example:”, (3) the EXAMPLE, and (4) HELP explaining how the example relates to the claim.

Now a most solemn piece of advice: INTEGRATE CLAIMS AND EXAMPLES. Some authors present a group of examples followed by various claims about the examples. Those claims may even be embedded in lengthy discussion.[2] Avoid this. Try to keep each example as close as possible to the claim it illustrates.

2.4 Enrichment

Typically, an example is a text fragment enriched in various ways. Some enrichments are strictly linguistic:

- The boundaries between morphemes are marked by hyphens.
- Each morpheme is identified by a gloss (“tag”).
- A free translation is included, usually in some major language. (For grammars to be delivered in two or more languages, a translation must be given in each of those languages. The software that renders the grammar would display the appropriate translation(s) based on user-defined preferences and the context.)
- Structural units may be indicated with brackets.
- Categories may be indicated, usually as subscripts to brackets.
- Functions (e.g., subject, instrument, source, and so forth) may be given for noun phrases.
- Grammaticality judgements may be indicated (*, **, ?, ??).
- Subscripted indices may indicate coreference or disjoint reference between referring expressions.
- Empty categories may be indicated: Ø, e, t, PRO, pro

- Ellipsis marking may indicate that the example is only part of a larger unit.
- Explanatory notes may be attached to morphemes or larger units. In an ink-on-paper environment these might be rendered as footnotes or endnotes. In an electronic delivery environment they might be activated by the user (by clicking or hovering with mouse).
- A digitized recording of the example might be attached to the example.

And so forth.

Some enrichments are aids to the reader, particularly to direct attention:

- Attention may be directed to a particular aspect by italics, bolding, underlining, or some other highlighting mechanism.
- Multiple examples may be conflated by means of braces or parentheses to facilitate the comparison between alternatives.
- Punctuation may be added.

And so forth.

Some enrichments serve to identify the example in the context of the document in which it is cited:

- Normally each example bears a number identifying it in the description. This number is used for referring to the example, either from within the description or from some other document.
- An example may employ internal identifiers for alternatives within the example itself; e.g., "See example 44b." might direct the reader to the second alternative of example 44 (on page 13).

And so forth.

Some enrichments give information about the example (its source, context, etc.):

- Speaker: name, age, sex, dialect,…
- Context: when, where, to whom,… the sentence was spoken.
- Register: formal, colloquial,…
- Mode of production: written, oral, recorded, videotaped,…
- Textual context: a reference to the text from which the example is drawn, and its position in that text.
- Situational context: the context in which an utterance might be used, as illustrated in (5):

(5) qam-pis maqa-ma-ška-nki { a. -mi (-DIR) b. -ši (-RPT) c. -či (-CNJ) }
you-ALSO hit-⇒1-PRF-2
'You also hit me.'

a. SITUATION: I felt you hit me and realized it was you.
b. SITUATION: I was unconscious when you hit me, but someone told me that you did so.
c. SITUATION: Various people hit me and I surmise that you were one of them.

- Residence: the document, archive, or collection, … possessing the text from which the example is taken

And so forth.

Elements of these various types are combined and given visual form, traditionally on a printed page, but now increasingly on a computer screen.

3. Layout

An example must be laid out, that is, its various parts must be located on the page or screen. This may involve the conflation of two or more examples, deciding how to wrap long lines or break examples from one page to the next, and adding enrichments, particularly those that direct the reader's attention. We discuss these in turn.

3.1 The layout of the major elements

There are only loose, unwritten conventions for how the parts of an example are laid out.[3] A rather standard form of example has a number (NUMBER); the text fragment, with hyphens dividing morphemes (MORPHEMES); morpheme glosses, usually aligned word-by-word but sometimes aligned morpheme-by-morpheme (GLOSSES); and a free translation ('TRANSLATION'). These are laid out as follows:

(NUMBER) MORPHEMES
GLOSSES
'TRANSLATION'

For example:

(6) [[yapya-y]-ta uša-na-n]-ta-ši šuya-ra-yka-n
plow-INF-OBJ finish-SUB-3P-OBJ-RPT wait-DUR-IMPFV-3
'He is waiting for him to finish plowing.'

In (6) the morphemes are represented phonemically with characters familiar to linguists. There are various reasons for also including the example written with the writing system used by speakers of the language:

- Linguistically-oriented representations (phonetic, phonemic, morphophonemic) may be inaccessible to speakers of the language whereas including the traditional/conventional writing system may make it easy for them to read.[4]
- Linguists who study a language seriously should learn its writing system so as to be able to benefit from other documents written in the language. This learning can occur simply by seeing the practical orthography along with a more linguistically-oriented representation.
- The practical orthography may have information not contained in the morphemic representation. For example, in example (10) below the practical orthography represents phonetic detail not indicated by the morphemic form.

The most normal place for the practical orthography (WRITING) is perhaps on the very first line. Because it is primarily for readers who can read and understand it, it is not necessary to align it with the glosses.

(NUMBER) WRITING
MORPHEMES
GLOSSES
'TRANSLATION'

For example:

(7) "¡Ama aywaychu!" nir willashcä.
ama aywa-y-ču ni-r wiλa-ška-:
no go-2IMP-NEG say-SS advise-PRF-1
'I told him not to go. (lit. I advised him saying "Do not go!")'

When examples are short, this sort of layout may waste space. When printed, it can increase the cost. When viewed on-line, it can push relevant text off the screen. Therefore, when space permits, it may be desirable to use alternative layouts. The translation, for example, might follow the morphemic representation:

(NUMBER) WRITING
MORPHEMES ‘TRANSLATION’
GLOSSES

For example:

(8) Liguiyta yachachimanga.
ligi-y-ta yača-či-ma-nqa ‘He will teach me to read.’
read-INF-OBJ learn-CAUS-⇒1-3FUT

Or the written form may also fit there:

(NUMBER) MORPHEMES WRITING ‘TRANSLATION’
GLOSSES

For example:

(9) huk runa ka-ša Juc runa casha. ‘There was a man.’
one man be-3PRF

Short examples might all fit on one line:

(NUMBER) MORPHEMES (GLOSSES) WRITING ‘TRANSLATION’

For example:

(10) a. imana-ša-taq (what.do-3PRF-¿?) ¿Imanashataj? ‘What did he do?’
b. imana-šaq-taq (what.do-1FUT-¿?) ¿Imanashätaj? ‘What will I do?’

If — as I am assuming — the layout of the major elements depends on the available space, and if in a web-based environment column width is under the control of the reader, then the rendering engine should include a component that adjusts the layout depending on the available space and user preferences.

3.2 Conflation

Two or more examples may be conflated by means of braces, parentheses or brackets. In some cases only words are conflated; in others the morphemes within a word might be conflated. For example, consider the following, taken from the *International Journal of American Linguistics* 65:159:

(44a) li:-ta-pa:-chi’:-ní:t tasiw caja
INSTR-INGR-belly-tie-PFV rope box
‘The box has been tied up with a rope.’

(44b)

li:-ta-maq-chi':-ní:t	tasiw	caja
INSTR-INGR-body-tie-PFV	rope	box

'The box has been tied up with a rope.'

These could have been conflated as follows:

(44)

li:-ta-	{ a. pa:	-chi':-ní:t	tasiw	caja
	belly			
INSTR-INGR	b. maq	-tie-PFV	rope	box
	body }			

'The box has been tied up with a rope.'

There are various reasons for conflating examples:

- It makes the example more readable: without conflation the reader must scan the examples to isolate the parts being compared or contrasted (as discussed regarding example (11) below). With conflation this is immediately obvious.
- It makes better use of space. Thus, for a printed page it is more economical, and for a computer monitor it allows more context to be kept in view.
- It may simplify wrapping examples across lines and breaking lines over pages: Without conflation, two or more parallel lines normally wrap or break independently, which means that as the column is narrowed the layout becomes increasingly difficult to read. When two or more examples are conflated to a single line, this is more likely to wrap or break without creating problems (assuming that the portion in braces moves as a piece).

Examples are conventionally conflated with either braces or brackets, as discussed in the following sections.

3.2.1 *Braces*

Linguistic Inquiry is now virtually devoid of braces except the *characters* { and }. (This is probably as the result of making the journal available on-line, thus submitting to the limitations of HTML.) This has a cost; for example, consider the difficulty of reading and the wasted space in (11), which is example 33 from *LI* 30:658. ("BP" stands for Brazilian Portuguese.)

(11) a. (BP)

Eu	encontrei	as	minhas	velhas	amigas	e
I	met	the.F.PL	my.F.PL	old.F.PL	friends.F.PL	and

amigos	juntos.
friends.M.PL	together.M.PL

b. Eu encontrei as minhas velhas amigas e
I met the.F.PL my.F.PL old.F.PL friends.F.PL and
amigos no mesomo dia.
friends.M.PL on.the same day.
'I met my famous old female friends and male friends together/on the same day.'

This could be conflated as follows, both making it easier to read and saving space:

(12) Eu encontrei as minhas velhas amigas e amigos (BP)
I met the.F.PL my.F.PL old.F.PL friends.F.PL and friends.M.PL
{ a. juntos.
together.M.PL
b. no mesomo dia.
on.the same day. } 'I met my famous old female friends and male friends { a. together.'
b. on the same day.' }

Here are some further examples (from Huallaga Quechua), ones that illustrate both the utility and potential complexities of using braces. In (13) note the subscripts *i* and *j* in both the morphemic representations and the translation:[5]

(13) a. Magarcamaptin jaytashurayqui.
maqa-rkU-ma- { a. -pti DS
b.*-špa SS } -n$_i$ hayta-šu-ra-yki$_j$
hit-UP-⇒1 3P kick-⇒2-PST-2P
a. 'After he$_i$ hit me, he$_j$ kicked you. (*i*≠*j*)'

There may be braces within braces, that is, conflation within conflation. (14) conflates four examples. ((14b) and (c) are grammatical while (14a) and (d) are not.)

(14) b. Magarcushpan jaytamaran.
c. Magarcur jaytamaran.
maqa-rku- { -špa SS { a.*-ø
b. -n (-3P) }
-r SS { c. -ø
d.*-nin (-3P) } } hayta-ma-ra-n
hit-ARR kick-⇒1-PST-3
b,c. 'After he$_i$ hit him$_j$, he$_i$ kicked me.'

Note that in (14) there are no right braces matching the smaller left braces. Is this good practice? Suppressing the right braces may look better, but it might complicate some computational tasks.

In (15) a single left brace is matched by two right braces. Again, is this good practice?

$$\text{(15)}\quad \begin{array}{ll}\text{pay-ta} & \text{rika-}\\ \text{him-OBJ} & \text{see}\end{array} \left\{\begin{array}{l} \left.\begin{array}{ll}\text{a.} & \text{-na-:-paq}\\ & \text{-SUB-1P-PUR}\\ \text{b.} & \text{-q}\\ & \text{-SUB}\end{array}\right\} \begin{array}{l}\textbf{šamu}\text{-ška-:}\\ \text{come-PRF-1}\end{array} \\ \left.\begin{array}{ll}\text{c.} & \text{-na-:-paq}\\ & \text{-SUB-1P-PUR}\\ \text{d.} & \text{*-q}\\ & \text{-SUB}\end{array}\right\} \begin{array}{l}\textbf{šuya-ra-yka}\text{-ška-:}\\ \text{wait-DUR-IMPFV-PRF-1}\end{array}\end{array}\right.$$

a,b. ‘I came to see him.’
c. ‘I was waiting to see him.’

The possibilities, of course, are limitless, and *this is a problem*! It would be nice to have a statement of “best practice” that would gently constrain authors’ inventiveness.

Authors must consider the cost of using conflation mechanisms, keeping in mind that readers’ familiarity with the use of braces, indices, and such devices. Use them judiciously. Above all, avoid needless complexity.

3.2.2 *Brackets*

Square brackets are sometimes used in contrast to (curly) braces to signal a correspondence among bracketed elements. For example, consider (16):

(16)	a.	**kay**-man	aywa-**mu**-n	‘He comes here.’
		here-GOAL	go-to.here-3	
	b.	**čay**-man	aywa-n	‘He goes there.’
		there-GOAL	go-3	

This might be conflated as in (17), indicating that *kay* co-occurs with *-mu* and *‘He comes here.’*, while *čay* co-occurs with the absence of *-mu* and *‘He goes there.’*:

$$\text{(17)}\quad \begin{bmatrix}\text{kay}\\ \text{here}\\ \text{čay}\\ \text{there}\end{bmatrix} \begin{array}{ll}\text{-man} & \text{aywa-}\\ \text{GOAL} & \text{go}\end{array} \begin{bmatrix}\text{- mu}\\ \text{TO.HERE}\\ \text{-ø}\end{bmatrix} \begin{array}{l}\text{-n}\\ \text{3}\end{array} \begin{bmatrix}\text{‘He comes here.’}\\ \text{‘He goes there.’}\end{bmatrix}$$

My impression is that square brackets are used less and less, and I would like to think that they are a thing of the past. Perhaps this is because it is possible to more explicitly express the correspondence of elements using internal identifiers like "a." and "b." as in (18):

(18) $$\left\{\begin{array}{l}\text{a. kay}\\ \quad\text{here}\\ \text{b. čay}\\ \quad\text{there}\end{array}\right\}\begin{array}{ll}\text{-man} & \text{aywa-}\\ \textsc{goal} & \text{go}\end{array}\left\{\begin{array}{l}\text{a. -mu}\\ \quad\textsc{to.here}\\ \text{b. -Ø}\end{array}\right\}\begin{array}{c}\text{-n}\\ \text{3}\end{array}\quad\left\{\begin{array}{l}\text{a. 'He comes here.'}\\ \text{b. 'He goes there.'}\end{array}\right\}$$

So I recommend NOT using square brackets. Maybe braces, but not brackets.

3.2.3 *In-line conflation*

There are in-line conflations:

- X A/B/… Y is equivalent to $\text{X}\left\{\begin{array}{c}\text{A}\\ \text{B}\\ \vdots\end{array}\right\}\text{Y}$, which conflates $\left\{\begin{array}{c}\text{X A Y}\\ \text{X B Y}\\ \vdots\end{array}\right\}$.

The examples in (19) are from *LI* 30:545:

(19) If/As/When you eat more, you want correspondingly less.
If/*As you had eaten more, you would want less.

- X (*A) Y is equivalent to $\left\{\begin{array}{c}\text{Ø}\\ \text{*A}\end{array}\right\}$, which conflates $\left\{\begin{array}{c}\text{*X Y}\\ \text{X A Y}\end{array}\right\}$.

Example (20) is from *LI* 30:568. ("t" is a trace.)

(20) This is the kind of rice that the quicker (*that) you cook t, the better it tastes.

- X *(A) Y is equivalent to $\left\{\begin{array}{c}\text{*Ø}\\ \text{A}\end{array}\right\}$, which conflates $\left\{\begin{array}{c}\text{*X Y}\\ \text{X A Y}\end{array}\right\}$.

Authors should assess the costs and benefits of such conflations for the intended audience. In language discriptions intended for non-technical audiences — both present and future — it may be wise to limit the use of conflation. (However, see Section 3.4 below.)

3.3 Line wrapping and page breaks

When an example must be broken across a page boundary, it is important that this be done at certain points and not at others. For example, the glosses should never be separated from the morphemes to which they correspond.

Likewise, when an example is too long to fit on a single line, it must be "wrapped" in a way that does not interpose text between, say, the morpheme decomposition and the corresponding glosses.

To break some lines attractively may require hyphenation. For example, for the Spanish version of my Huallaga grammar, both the Quechua written form (practical orthography) and the Spanish translation were hyphenated, that is, "discretionary hyphens" were computationally introduced. Note: hyphenation differs from language to language subject to convention, syllable structure, and even subjective esthetic criteria.

Although quite obvious, we should not forget that the space in which an example is rendered depends on the document context: if it is embedded within an item in a list, where each item is indented, then the effective column width for the example is correspondingly narrower.

3.4 Some future possiblities

Documents are increasingly published electronically and read on screens rather from printed pages. Technology will progressively enhance the display of documents in ways that are not possible with ink-on-paper delivery. Software could be developed to enhance the presentation of language descriptions; here are some possibilities:

inspect the context: Traditionally, what you see is all you get. Although an example might be a fragment of a text, when used in a linguistic description, the reader can not see what preceeds or follows it in the original text. In the future, when examples are fragments of online texts, software should allow the user to dynamically inspect the original text surrounding the example.

toggle on/off parts: It may be useful to turn off or on the display of certain kinds of information. For example, native speakers may wish to toggle off parts they do not need, such as the morphemic representation, glosses and translation. Linguists not familiar with the language may wish to toggle off the practical orthography, while linguists familiar with the language may wish to suppress the gloss. Users should be able to tailor the display of information to meet their needs and preferences.

buttons and hot zones: Buttons could be provided to activate certain kinds of secondary information, e.g., the speaker's biographical information, the context of use, the example's "residence," and so forth. Perhaps if the gloss is toggled off, morphemes could be "hot," so that clicking on or hovering over them would trigger the display of information about the morpheme: the gloss, the category, perhaps even a lexical entry for that morpheme.

enhanced focus mechanisms: Traditionally attention is directed by static effects like bold or italic type, or by underlining. Now it should be possible to use coloring and effects like blinking. It might be useful to have three variants of comparison, one to signal 'note the similarity of these', one to signal 'note difference between these', and a default for simple comparison.

control conflation It should be possible to control conflation, with the default appearance determined partly by the author and partly by the reader. For example, the author may give a conflated form, but a reader may wish to "deconflate" the alternatives to see them as a list of sentences without braces, brackets or parentheses.

At present electronic documentents can be enhanced in ways like those just mentioned only by people with considerable technical training and skill. Linguists lack the software with which to implement such possibilities in the course of writing a language description. Ideally language data would be managed — and grammars written — in a computational framework that integrates grammar and corpus, with examples existing in the corpus but accessed from the grammar. Examples would not be "taken from" a text but displayed therein.

Grammar writers need hospitable authoring environments, with tools that are powerful and flexible, yet reasonably easy to learn and use. Until these are available we labor under the limitations of ink-on-paper.

Notes

[*] This paper draws from a paper presented at the Workshop on Web-Based Language Documentation and Description, December 2000, Philadelphia. That paper is available at www.ldc.upenn.edu/exploration/expl2000/papers/weber/weber.pdf.

[1] My grammar of Huallaga Quechua, for example, has over 1700 examples.

[2] Chomsky's *Lectures on Government and Binding* has many fine examples! This may be appropriate in a context for which theory is primary, with examples simply providing grist for the theoretical mill. It is quite a different matter for a language description, in which theory is generally a servant to description rather than its master.

3. We need guidelines leading to good practice and curbing individuals' tendencies toward the idiosyncratic. Grammar writers need a style sheet for examples!

4. Except for material dealing with phonology, English examples use the practical orthography. Readers would be very put off if they had to read English examples in a phonetic, phonemic, or morphophonemic representation.

5. Avoid examples like (13) and (14), obviously elicited and so richly endowed with violence.

Author's address:

David J. Weber
7264 W. Main St.
Westmoreland, NY 13490
david_weber@sil.org

AnnCorra : Annotating Corpora
Guidelines For POS And Chunk Annotation For Indian Languages

Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, Rajeev Sangal
Language Technologies Research Centre
IIIT, Hyderabad

Date : 15-12-2006

OUTLINE

1. Introduction

The significance of large annotated corpora in the present day NLP is widely known. Annotated corpora serve as an important tool for investigators of natural language processing, speech recognition and other related areas. It proves to be a basic building block for constructing statistical models for automatic processing of natural languages.

Many such corpora are available for languages across the world and have proved to be a useful step towards natural language processing. Coming to the scenario for Indian languages, not much work has been carried out on the front of automatic processing of Hindi or any other Indian language. The main bottleneck being unavailability of an annotated corpora, large enough to experiment statistical algorithms.

Annotation of corpora (AnnCorra) can be done at various levels viz, part of speech, phrase/clause level, dependency level, etc. Part of speech tagging forms the basic step towards building an annotated corpus. Chunking can form the next level of tagging.

The task of annotating corpora of several Indian languages has been taken up in the Indian Language Machine Translation (ILMT) project. ILMT is a project in which a number of institutes have come together to form a consortium and work towards developing MT systems for various Indian language pairs.

A primary requirement of such an effort is defining standards for various sub tasks. Thus, standardization of annotation schemes for various annotation tasks becomes a crucial step in this direction.

The issues related to defining standards for POS/Chunk tagging schemes were discussed by scholars from various Indian institutes by way of holding meetings etc. and some standards have been arrived at.

2. Objective

The purpose of the meetings was to arrive at standard tagging scheme for POS tagging and chunking for annotating Indian languages (AnnCorra) and come up with the tags which are exhaustive for the task of annotation for a larger group of languages, specially, Indian languages. The present document gives a detailed description of the tags which have been defined for the tagging schemes and elaborates the motivations behind the selection of these tags. The document also discusses various issues that were addressed while preparing the tag sets and how they have been resolved.

3. Some Assumptions

3.1 During the workshop it was decided to base the discussion and decisions about various tags on the following basic assumptions which everybody agreed on :

i) The tags should be common for all Indian languages
ii) It should be comprehensive/ complete
iii) It should be simple. Maintaining simplicity is important for the following two reasons :

(a) Ease of Learning

(b) Consistency in annotation

3.2 Another important point which was discussed and agreed upon was that POS tagging is NOT a replacement for morph analyser. A 'word' in a text carries the following linguistic knowledge

a) grammatical category and
b) grammatical features such as gender, number, person etc. The POS tag should be based on the 'category' of the word and the features can be acquired from the morph analyser.

4. Issues in Tag Set Design

This section deals with some of the issues related to any POS tagger and the policy that we have adopted to deal with each of these issues for our purpose.

The first step towards developing POS annotated corpus is to come up with an appropriate tags. The major issues that need to be resolved at this stage are :

1. Fineness vs Coarseness in linguistic analysis
2. Syntactic Function vs lexical category
3. New tags vs tags close to existing English tags

4.1 Fineness vs Coarseness

An issue which always comes up while deciding tags for the annotation task is whether the tags should capture 'fine grained' linguistic knowledge or keep it 'coarse'. In other words, a decision has to be taken whether or not the tags will account for finer distinctions of the parts of speech features. For example, it has to be decided if plurality, gender and other such information will be marked distinctly or only the lexical category of a given word should be marked.

It was decided to come up with a set of tags which avoids 'finer' distinctions. The motivation behind this is to have less number of tags since less number of tags lead to efficient machine learning. Further, accuracy of manual tagging is higher when the number of tags is less.

However, an issue of general concern is that in an effort to reduce the number of tags we should not miss out on crucial information related to grammatical and other relevant linguistic knowledge which is encoded in a word, particularly in agglutinating languages, eg, Tamil, Telugu and many other Indian languages. If tags are too coarse, some crucial information for further processing might be missed out. As mentioned above, primarily the required knowledge for a given lexical item is its grammatical category, the features specifying its grammatical information and any other information suffixed into it. For example,

Telugu word ' *rAmudA* (Is it Ram ?)' contains the following information <category (noun)+grammatical features(masculine, singular) + question>. The word by itself is a bundle of linguistic information. Morph analyser provides all the knowledge that is contained in a word. It was decided that any linguistic knowledge that can be acquired from any other source (such as morph analyser) need not be incorporated in the POS. As mentioned above, POS tagger is not a replacement for morph analyser. In fact, features from morph analyser can be used for enhancing the performance of a POS tagger. The additional knowledge of a POS given by a POS tagger can be used to disambiguate the multiple answers provided by a morph analyser.

On the other hand, we agree that too coarse an analysis is not of much use. Essentially, we need to strike a balance between fineness and coarseness. The analysis should not be so fine as to hamper machine learning and also should not be so coarse as to miss out important information. It is also felt that fine distinctions are not relevant for many of the applications(like sentence level parsing, dependency marking, etc.) for which the tagger may be used in future.

However, it is well understood that plurality and other such information is crucial if the POS tagged corpora is used for any application which needs the agreement information. In case such information is needed at a later stage, the same tag set can be extended to encompass information such as plurality etc as well. This can be done by providing certain heuristics or linguistic rules.

Thus, to begin with, it has been decided to adopt a coarse part of speech analysis. At the same time, wherever it is found essential, finer analysis is incorporated. Also, there is a basic understanding that wherever/whenever essential, the tags containing finer linguistic knowledge can be incorporated. An example of where finer analysis becomes crucial has been given below. Take the Hindi sentence (h1) below :

h1. *AsamAna*_NN *meM*_PSP *uDane*_VM *vAlA*_PSP *ghoDA*_NN
'sky' 'in' 'flying' 'horse
*nIce*_NST *utara*_VM *AyA*_VAUX.
'down' “descend” “came”

In (h1) above *uDane* is a noun derived from a verb. The word *AsamAna* is an argument of *uDane* and not of *'nIce utara AyA* – another verb in the sentence. It is crucial to retain the information that *uDane,* though functioning as a noun now, is derived from a verb and can take its own arguments. In order to preserve such crucial information a finer analysis is essential. Therefore, a distinct tag needs to be introduced for such expressions. In the current tagging scheme *uDane* will be annotated as a 'main verb (VM)' at the POS level. However, the information that it is functioning like a noun will be   captured at the chunking level by introducing a distinct chunk tag VGNN (discussed in details under Section III on Chunking).

4.2 Syntactic Function vs Lexical Category

A word belonging to a particular lexical category may function differently in a given context. For example, the lexical category of *harijana* in Hindi is a noun . However, functionally, *harijana* is used as an adjective in (h2) below,

h2. *eka dina pAzca baje khabara AyI ki koI* ***harijana***
'one' 'day' 'five' 'o'clock' 'news' 'came' 'that' 'some' 'harijana'
bAlaka *unase milanA cAhatA hE*

'young boy' 'him' 'to meet' 'wants' 'is'
"One day, a message came at five o'clock that some 'harijana' boy wanted to meet him".

Such cases require a decision on whether to tag a word according to its lexical category or by its syntactic category. Since the word in a context has syntactic relevance, it appears natural to tag it based on its syntactic information. However, such a decision may lead to further complications.

In AnnCorra, the syntactic function of a word is not considered for POS tagging. Since the word is always tagged according to its lexical category there is consistency in tagging. This reduces confusion involved in manual tagging. Also the machine is able to establish a word-tag relation which leads to efficient machine learning.

In short, it was decided that syntactic and semantic/pragmatic functions were not to be the basis of deciding a POS tag.

4.3. New Tags vs Tags from a Standard Tagger

Another point that was considered while deciding the tags was whether to come up with a totally new tag set or take any other standard tagger as a reference and make modifications in it according to the objective of the new tagger. It was felt that the later option is often better because the tag names which are assigned by an existing tagger may be familiar to the users and thus can be easier to adopt for a new language rather than a totally new one. It saves time in getting familiar to the new tags and then work on it.

The Penn tags are most commonly used tags for English. Many tag sets designed subsequently have been a variant of this tag set (eg. Lancaster tag set). So, while deciding the tags for this tagger, the Penn tags have been used as a benchmark. Since the Penn tag set is an established tag set for English, we have used the same tags as the Penn tags for common lexical types. However, new tags have been introduced wherever Penn tags have been found inadequate for Indian language descriptions. For example, for verbs none of the Penn tags have been used. Instead, AnnCorra has only two tags for annotating verbs, VM (main verb) and VAUX (auxiliary verb).

5. POS tags Chosen for the Current Scheme

This section gives the rationale behind each tag that has been chosen in this tag set.

5.1.1 NN Noun

The tag NN for nouns has been adopted from Penn tags as such. The Penn tag set makes a distinction between noun singular (NN) and noun plural (NNS). As mentioned earlier, distinct tags based on grammatical information are avoided in IL tagging scheme. Any information that can be obtained from any other source is not incorporated in the POS tag. Plurality, for example, can be obtained from a morph analyzer. Moreover, as mentioned earlier, if a particular information is considered crucial at the POS tagging level itself, it can be incorporated at a later date with the help of heuristics and linguistic rules. This approach brings the number of tags down, and helps achieve simplicity, consistency, better machine learning with a small corpora etc. Therefore, the current scheme has only one tag (NN) for common nouns without getting into any distinction based on the grammatical information contained in a given noun word

5.1.2 NST Noun denoting spatial and temporal expressions

A tag NST has been included to cover an important phenomenon of Indian languages. Certain expressions such as '*Upara*' (above/up), '*nIce*' (below) '*pahale*' (before), 'Age' (front) etc are content words denoting time and space. These expressions, however, are used in various ways. For example,

5.1.2.1 These words often occur as temporal or spatial arguments of a verb in a given sentence taking the appropriate *vibhakti* (case marker):

h3. *vaha* ***Upara*** *so rahA thA* .
'he' 'upstairs' 'sleep' 'PROG' 'was'
“He was sleepign upstairs”.

h4. *vaha* ***pahale se*** *kamare meM bEThA thA* .
'he' 'beforehand' 'from' ' room' 'in' 'sitting' 'was'
“He was sitting in the room from beforehand”

h5. *tuma* ***bAhara*** *bETho*
'you' 'outside' 'sit'
“You sit outside”.

Apart from functioning like an argument of a verb, these elements also modify another noun taking postposition 'kA'.

h6. *usakA baDZA bhAI* ***Upara ke hisse meM*** *rahatA hE*
'his' 'elder' 'brother' 'upstairs' 'of' 'portion' 'in' 'live' 'PRES'
“His elder brother lives in the upper portion of the house”.

5.1.2.2 Apart from occuring as a nominal expression, they also occur as a part of a postposition along with 'ke'. For example,

h7. *ghaDZe* ***ke Upara*** *thAlI rakhI hE.*
'pot' 'of' 'above' 'plate' 'kept' 'is'
The plate is kept on the pot”.

h8. *tuma ghara ke* **bAhara** *bETho*
'you' 'home' 'of' 'outside' 'sit'
"You sit outside the house".

'Upara' and '*bAhara*' are parts of complex postpositions *'ke Upara'* and '*ke bAhara*' in (h6) and (h7) respectively which can be translated into English prepositions 'on' and 'outside'.

For tagging such words, one possible option is to tag them according to their syntactic function in the given context. For example in 5.2.2 (h7) above, the word '*Upara*' is occurring as part of a postposition or a relation marker. It can, therefore, be marked as a postposition. Similarly, in 5.2.1. (h3) and (h6) above, it is a noun, therefore, mark it as a noun and so on. Alternatively, since these words are more like nouns, as is evident from 5.2.1 above they can be tagged as nouns in all there occurrences. The same would apply to '*bAhAra'* (outside) in examples examples (h4), (h5) and (h8).
However, if we follow any of the above approaches we miss out on the fact that this class of words is slightly different from other nouns. These are nouns which indicate 'location' or 'time'. At the same time, they also function as postpositions in certain contexts. Moreover, such words, if tagged according to their syntactic function, will hamper machine learning. Considering their special status, it was considered whether to introduce a new tag, NST, for such expressions. The following five possibilities were discussed :

a) Tag both (h5) & (h8) as NN
b) Tag both (h5) & (h8) as NST
c) Tag (h5) as NN & (h8) as NST
d) Tag (h5) as NST & (h8) as PSP
e) Tag (h5) as NN & (h8) as PSP

After considering all the above, the decision was taken in favour of (b). The decision was primarily based on the following observations:

(i) *'bAhara'* in both (h5) and (h8) denotes the same expression (place expression 'outside')
(ii) In both (h5) and (h8), *'bAhara'* can take a vibhakti like a noun (**bAhara ko** bETho, ghara **ke bAhara ko** bETho)
(iii) If a single tag is kept for both the usages, the decision making for annotators would also be easier.

Therefore, a new tag **NST** is introduced for such expressions. The tag **NST** will be used for a finite set of such words in any language. For example, Hindi has *Age* (front), *pIche* (behind), *Upara* (above/upstairs), *nIce* (below/down), *bAda* (after), *pahale* (before), *andara* (inside), *bAhara (outside)* etc.

5.2 NNP Proper Nouns

The need for a separate tag for proper nouns and its usability was discussed. Following points were raised against the inclusion of a separate tag for proper nouns :

a) Indian languages, unlike English, do not have any specific marker for proper nouns in orthographic conventions. English proper nouns begin with a capital letter which distinguishes them from common nouns.
b) All the words which occur as proper nouns in Indian languages can also occur as common nouns denoting a lexical meaning. For example,
English : John, Harry, Mary occur only as proper nouns whereas
Hindi : *aTala bihArI, saritA, aravinda* etc are used as 'names' and they also belong to grammatical categories of words with various senses . For example given below is a list of Hindi words with their grammatical class and sense.

aTala	adj	immovable
bihArI	adj	from Bihar
saritA	noun	river
aravinda	noun	lotus

Any of the above words can occur in texts as common lexical items or as proper names. (h9) - (h11) below show their occurrences as proper nouns,

h9. ***atala bihAri bAjapaI*** *bhArata ke pradhAna mantrI the.*
'Atal' 'Bihari' 'Vajpayee' 'India' 'of' 'prime' 'minister' 'was'
"Atal Behari Vajpayee was the Prime Minister of India".

h10. merI mitra **saritA** tAIvAna jA rahI hE.
'my' 'friend' 'Sarita' 'Taiwan' 'go' 'PROG' 'is'
"My friend Sarita is going to Taiwan"

h11. ***aravinda*** *ne mohana ko kitAba dI.*
'Aravind' 'erg' 'Mohan' 'to' 'book' 'gave'
"Aravind gave the book to Mohan".

Therefore, in the Indian languages' context, annotating proper nouns with a separate tag will not be very fruitful from machine learning point of view. In fact, the identification of proper nouns can be better achieved by named entity filters.

Another point that was considered in this context was the effort involved in manual tagging of proper nouns in a given text. It is felt that not much extra effort is required in manual tagging of proper nouns. However, the data annotated with proper nouns can be useful for certain applications. Therefore, there is no harm in marking the information if it does not require much effort.

Finally, it was decided to have a separate tag for proper nouns for manual annotation and ignore it for machine learning algorithms. Following this decision, the tag **NNP** is included in the tag set. This tag is the same as the

Penn tag for proper nouns. However, in this case also AnnCorra has only one tag for both singular and plural proper nouns unlike Penn tags where a distinction is made between proper noun singular and proper noun plural by having two tags NNP and NNPS respectively.

5.3.1 PRP Pronoun

Penn tags make a distinction between personal pronouns and possessive pronouns. This distinction is avoided here. All pronouns are marked as PRP. In Indian languages all pronouns inflect for all cases (accusative, dative, possessive etc.). In case we have a separate tag for possessive pronouns, new tags will have to be designed for all the other cases as well. This will increase the number of tags which is unnecessary. So only one tag is used for all the pronouns. The necessity for keeping a separate tag for pronouns was also discussed, as linguistically, a pronoun is a variable and functionally it is a noun. However, it was decided that the tag for pronouns will be helpful for anaphora resolution tasks and should be retained.

5.3.2 DEM Demonstratives

The tag 'DEM' has been included to mark demonstratives. The necessity of including a tag for demonstratives was felt to cover the distinction between a pronoun and a demonstrative. For example,

h12. ***vaha ladakA*** *merA bhAI* *hE* (demnostrative)
'that' 'boy' 'my' 'brother' 'is'
h13. ***vaha*** *merA bhAI* *hE* (pronoun)
'he' 'my' 'brother' 'is'

Many Indian languages have different words for demonstrative adjectives and pronouns. A better evidence for including a separate tag for demonstratives is from the following Telugu examples,

t1. *A abbAyi nA tammudu*
'that' 'boy' 'my' 'brother'
t2. *atanu nA tammudu*
'he' 'my' 'brother'
(Telugu does not have a copula 'be' in the present tense)

5.4 VM Verb Main

Verbal constructions in languages may be composed of more than one word sequences. Typically, a verb group sequence contains a main verb and one more auxiliaries (V AUX AUX ... ...). In the current tagging scheme the support verbs (such as *dAlanA* in *kara dAlAtA hE*, *uThanA* in *cOMka uThA thA* etc) are also tagged as VAUX. The group can be finite or non-finite. The main verb need not be marked for finiteness. Normally, one of the auxiliaries carries the finiteness feature.

The necessity of marking the finiteness or non-finiteness in a verb was discussed extensively and everybody agreed that it was crucial to mark the distinction. However, languages such as Hindi, which have auxiliaries for marking tense, aspect and modalities pose a problem. The finiteness of a verbal expression is known only when we reach the last auxiliary of a verb group. Main verb of a finite verb group (leaving out the single word verbal expressions of the finite type – eg *vaha dillI gayA*) does not contain finiteness information. For example,

h14. *laDZakA seba* ***khAtA*** ***raHA wA***
'boy' 'apple' 'eating' 'PROG' 'was'
The boy had kept eating.

h15. *seba* ***khAtA*** ***huA*** *laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'
The boy eating the apple was going.

The expression *khAtA raHA* in (h29) above is finite and *khAtA huA* in (h3) is non finite. However, the main verb '*khAtA*' is non-finite in both the cases.

So, the issue is - whether to (1a) mark finiteness in *“**khAtA rahA thA (*** had kept eating*)*” at the lexical level on the main verb (khA) or (1b) on the auxiliary containing finiteness (wA) or (2) not mark it at the lexical level at all. All the three possibilities were discussed;
1) Mark the finiteness at the lexical level.

If we mark it at the lexical level, following possibilities are available :

1a) Mark the finiteness on the main verb, even though we know that the lexical item itself is not finite.

In this case, the annotator interprets the finiteness from the context. (The POS tags VF, VNF and VNN were earlier decided based on this approach). The main verb, therefore, is marked as finite consciously with a view that the group contains a 'verb root' and its auxiliaries (as TAM etc) is finite even though the main verb does not carry the finiteness at the lexical level. Although, this approach facilitates annotation of both the main verb and the finiteness (of the group) by a single tag, it allows tagging a lexical item (main verb) with the finiteness feature which it does not actually carry. So, this is not a neat solution.

1b) The second possibility is, mark the finiteness on the last auxiliary of the sequence. Here again the decision has to be taken from the context. This possibility was not considered since this also involves marking the verb finiteness at the lexical level.

2) Don't mark the finiteness at the lexical level. Instead mark it as indicated in (2a) or (2b) below.

2a) Introduce a new layer which groups the verb group and mark the verb group as finite or non-finite. This approach proposes the following :

(i) Annotate the main verb as **VM** (introduce a new tag). Thus,

h14a. *laDZakA seba* ***khAtA*_VM** *raHA thA*
'boy' 'apple' 'eating' 'PROG' 'was'

h15a. *seba* ***khAtA*_VM** *huA laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

(ii) Annotate the auxiliaries as **VAUX**,

h14a. *laDZakA seba khAtA*_VM *raHA*_**VAUX** *thA*_**VAUX**
'boy' 'apple' 'eating' 'PROG' 'was'
h15a. *seba khAtA*_VM *huA*_**VAUX** *laDZakA jA rahA thA*
'apple' 'eating' ' PROG' 'boy' ' go' 'PROG' 'was'

(iii) Group the verb group (before chunking) and annotate it as finite or non-finite as the case may be,

h14a. *laDZakA seba* [*khAtA*_VM *raHA*_VAUX *wA*_VAUX**]_VF**
'boy' 'apple' 'eating' 'PROG' 'was'
h15a. *seba* [*khAtA*_VM *huA*_VAUX**]_VNF** *laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

This approach is more faithful to the available linguistic information. However, it requires introducing another layer. So, this was not considered useful.

2b) Mark the finiteness at the chunk level,

In this approach, the lexical items are marked as in (2). No new layer is introduced. Instead, the decision is postponed to the chunk level. Since the finiteness is in the group, it is marked at the chunk level. This offers the best solution as it facilitates marking the linguistic information as it is without having to introduce a new layer.

h14a. *laDZakA seba* ((*khAtA*_VM *raHA*_VAUX *wA*_VAUX))**_VGF**
'boy' 'apple' 'eating' 'PROG' 'was'

h15a. *seba* ((*khAtA*_VM *huA*_VAUX))**_VGNF** *laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

In this case also the decision is made by looking at the entire group. (2b) was most preferred as it facilitates marking the linguistic information correctly, at the same time no new layer needs to be introduced. Therefore, the current tagging scheme has adopted this approach. Thus, the main verbs in a given verb group will be marked as **VM,** irrespective of whether the total verb group is finite of non finite. Given underneath are some examples of other verb group types :

1) **Non finite verb groups -** Non-finite verb groups can have two functions :

a) Adverbial participial, for example : *khAte-khAte* in the following Hindi sentence,

h16. *mEMne* ***khAte – khAte*** *ghode ko dekhA*
'I erg' 'while eating' 'horse' 'acc' 'saw'
"I saw a horse while eating".

The main verb in (h16) would be annotated as follows :

h16a. *mEMne* ***khAte – khAte*****_VM** *ghode ko dekhA*

b) Adjectival participial, for example : '*khAte Hue*' in the following Hindi sentence ,

h17. *mEMne ghAsa* ***khAte_VM hue*** *ghoDe ko dekhA* *
'I erg' 'grass' 'eating' 'PROG' 'horse' 'acc' 'saw'
I saw the horse eating grass.

(* (h17) is ambiguous in Hindi. The other sense that it can have is, *I saw the horse while (I was) eating grass*. In such cases, the annotator would disambiguate the sentence depending on the context and mark accordingly.)

2) Gerunds

Functionally, gerunds are nominals. However, even though they function like nouns, they are capable of taking their own arguments,eg. *pInA* in the following Hindi sentence can occur on its own or take an argument (given in parenthesis):

h18. (*sharAba)* ***pInA*****_VM** *sehata ke liye hAnikAraka hE.*
'liquor' 'drinking' 'health' 'for' 'harmful' 'is'
"Drinking (liquor) is bad for health"

h19. *mujhe* ***khAnA*****_VM** *acchA lagatA hai*
'to me' 'eating' 'good' 'appeals'
"I like eating"

h20. ***sunane meM*** *saba kuccha acchA lagatA hE* 'listening' 'in' 'all' 'things' 'good' 'appeal' 'is'

As mentioned above, noun '*sharAba*' in (h18) is an object of the verb '*pInA*' and has no relation to the main verb (*hE*). In order to be able to show the exact verb-argument structure in the sentence, it is essential that the crucial information of a noun derived from a verb is preserved. Therefore, even gerunds have to be marked as verbs. It is proposed that in keeping with the approach adopted for non-finite verbs, mark gerunds also as **VM** at the lexical level. For capturing the information that they are gerunds, such verbs will be marked as **VGNN** (see the section on Chunk tags for details) at the chunk level to capture their gerundial nature. The verbs having 'vAlA' vibhakti will also be marked as VM. For example, '*khonevAlA*' (one who looses).

5.5 VAUX Verb Auxiliary

All auxiliary verbs will be marked as VAUX. This tag has been adopted as such from the Penn tags. (For examples, see h14 – h16 above).

5.6 JJ Adjective

This tag is also taken from Penn tags. Penn tag set also makes a distinction between comparative and superlative adjectives. This has not been considered here. Therefore, in the current scheme for Indian languages, the tag JJ includes the 'tara' (comparative) and the 'tama' (superlative) forms of adjectives as well.
For example, Hindi *adhikatara* (more times), *sarvottama* (best), etc. will also be marked as JJ.

5.7 RB Adverb

For the adverbs also, the tag RB has been borrowed from Penn tags. Similar to the adjectives, Penn tags make a distinction between comparative and superlative adverbs as well. This distinction is not made in this tagger. This is in accordance with our philosophy of coarseness in linguistic analysis.
Another important decision for the use of RB for adverbs in the current scheme is that :-

(a) The tag RB will be used ONLY for 'manner adverbs' . Example,

h21. *vaha* *jaldI jaldI* *khA* *rahA* *thA*
'he' 'hurriedly' 'eat' 'PROG' 'was'

(b) The tag RB will NOT be used for the time and manner expressions unlike English where time and place expressions are also marked as RB. In our scheme, the time and manner expressions such as '*yahAz – vahAz, aba – waba* ' etc will be marked as PRP.

5.8 PSP Postposition

All Indian languages have the phenomenon of postpositions. Postpositions express certain grammatical functions such as case etc. The postposition will be marked as PSP in the current tagging scheme. For example,

h22. m*ohana kheta* ***meM*** *khAda dAla rahA thA*
'Mohan' 'field' 'in' 'fertilizer' 'put ' 'PROG' 'was'

meM in the above example is a postposition and will be tagged as PSP.
A postposition will be annotated as PSP ONLY if it is written separately. In case it is conjoined with the preceding word it will not be marked separately. For example, in Hindi pronouns the postpositions are conjoined with the pronoun,

h23. *mE**ne** usa**ko** bAzAra* ***meM*** *dekhA*
'I' 'him' 'market' 'in' 'saw'

(h23) above has three instances of 'postposition' (in bold) usage. The postpositions '*ne*' and '*ko*' are conjoined with the pronouns *mEM* and *usa* respectively. The third postposition '*meM*' is written separately. In the first two instances, the postposition will not be annotated. Such words will be annotated with the category of the head word. Therefore, the three instances mentioned above will be annotated as shown in (h23a) below :

h23a. *mE**ne***_**PRP** *usa**ko***_**PRP** *bAzAra*_**NN** ***meM*_PSP** *dekhA*

5.9 RP Particle

Expressions such as *bhI, to, jI, sA, hI, nA,* etc in Hindi would be marked as RP. The *nA* in the above list is different from the negative *nA.* Hindi and some other Indian languages have an ambiguous 'nA' which is used both for negation (NEG) and for reaffirmation (RP). Similarly, the particle *wo* is different from CC *wo*. For example in Bangla and Hindi:

Bangla : (b1) *tumi* ***nA*_RP** *khub dushtu*
'you' 'particle' 'very' 'naughty'
“You are very naughty” (comment)

Hindi : (h24) *tuma* ***nA*_RP**, *bahuta dushta ho*
'you' 'particle very naughty
“You are very naughty” (comment)

Bangla : (b2) *cheleta dushtu* ***nA*_NEG**
'the boy' 'naughty' 'not'
“The boy is not naughty”

Hindi : (h25) *mEM* ***nA*_NEG** *jA sakUMgA*
'I' 'not' 'go' 'will able'
“I will not be able to go”

Bangla : (b3) *binu yYoxi khAya* **to_CC** *Ami khAba*
'Binu' 'if' 'eats' 'then' 'I' 'will eat'
“If Binu eats then I will eat (too)”

Hindi : (h26) *yadi binu khAyegA* **wo_CC** *mEM khAUMgI*
'if' 'Binu' 'eats' 'then' 'I' 'will eat'
“Only if Binu eats, I will eat (too)”

Bangla : (b4) *Ami* ***to*****_RP** *jAni nA*
'I' 'particile' 'know' 'not'
“I don’t know”

Hindi : (h27) *mujhako* ***to*****_RP** *nahIM patA*
'I' 'particile' 'not' 'know'
“I don’t know”

5.10 CC Conjuncts(co-ordinating and subordinating)

The tag CC will be used for both, co-ordinating and subordinating conjuncts. The Penn tag set has used IN tag for prepositions and subordinating conjuncts. Their rationale behind this is that subordinating conjuncts and prepositions can be distinguished because subordinating conjuncts are followed by a clause and prepositions by a noun phrase.

But in the current tagger all connectives, other than prepositions, will be marked as CC.

h28. mohana bAzAra jA rahA hE **Ora_CC** ravi skUla jA rahA hE
'Mohan' 'market' 'go' 'PROG' 'is' 'and' 'Ravi' 'school' 'go' 'PROG' 'is'
“Mohan is going to the market and Ravi is going to the school”

h29. mohana ne mujhe batAyA **ki_CC** Aja bAzAra banda hE
'Mohan' 'erg' 'to me' 'told' 'that' 'today' 'market' 'close' 'is'
“Mohan told me that the market is closed today.”

5.11 WQ Question Words

The Penn tag set makes a distinction between various uses of 'wh-' words and marks them accordingly (WDT, WRB, WP, WQ etc). The 'wh-' words in English can act as questions, as relative pronouns and as determiners. However, for Indian languages we need not keep this distinction. Therefore, we tag the question words as WQ.

h30. ***kOna*** *AyA hE* ?
'who' 'come' 'has'
“Who has come ?

h31. *tuma kala* ***kyA*** *kara rahe ho* ?
'you' 'tomorrow' what' 'doing' 'are'
What are you doing tomorrow ?

h32. *tuma kala* ***kahAz*** *jA rahe ho* ?
'you' 'tomorrow' 'where' 'going' 'are'
"Where are you going tomorrow ?

h33. ***kyA*** *tuma kala Aoge* ?
'?' 'you' 'tomorrow' 'will come'
"Will you come tomorrow ?

5.12.1 QF Quantifiers

All quantifiers like Hindi *kama* (less), *jyAdA* (more), *bahuwa* (lots), etc. will be marked as QF.

h34. *vahAz* ***bahuta*_QF** *loga Aye the*
'there' many' 'people' 'came' 'was'
"Many people came there".

In case these words are used in constructions like '***baHutoM ne*** *jAne se inkAra kiyA*' ('many' 'by' 'to go' 'refused'; Many refues to go) where it is functioning like a noun, it will be marked as NN (noun). Quantifiers of number will be marked as below.

5.12.2 QC Cardinals

Any word denoting a cardinal number will be tagged as QC. Penn tag set has a tag CD for cardinal numbers and they have not talked of ordinals. For example,
h35. *vahAz* ***tIna*_QC** *loga bEThe the*
'there' 'three' 'people' 'sitting' 'were'
"Three people were sitting there"

5.12.3 QO Ordinals

Expressions denoting ordinals will be marked as QO.

h36. *mEMne kitAba* ***tIsare*_QO** *laDake ko dI thI*
'I' 'book' 'third' 'boy' 'to' 'give' 'was'
I gave the book to the third boy"

5.12.4 CL Classifiers

The tag CL has been included to mark classifiers. Many Indian languages have a rich classifier system. "A **classifier**, in linguistics, is a word or morpheme used in some languages to classify a noun according to its meaning" (http://en.wikipedia.org/wiki/Classifier_%28linguistics%29).

For example,

Telugu : (t2) *padi* ***mandi*** *pillalu* 'ten' 'persons' 'children'

Tamil : (tm1) *pattu* ***pEr*** *mANavarakaLa*
'ten' 'person' 'students'

The words 'mandi' (Telugu) and 'per' (Tamil) are classifiers which occur with numerals with human nouns. Such expressions when occurring separately (not suffixed with the noun) will be marked as CL. Therefore :

Telugu : (t2) *padi* ***mandi*_CL** *pillalu*
'ten' 'persons' 'children'

Tamil : (tm1) *pattu* ***pEr*_CL** *mANavarakaLa*
'ten' 'person' 'students'

5.13 INTF Intensifier

This tag is not present in Penn tag set. Words like *'bahuta*', *'kama'*, etc. when intensifying adjectives or adverbs will be annotated as INTF. Example,

h37. *hEdarAbAda meM aMgUra* **bahuta_INTF** *acche milate hEM*
'HyderabAd' 'in' 'grapes' 'very' 'good' 'available' 'are'
“Very good grapes are available in Hyderabad”.

5.14 INJ Interjection

The interjections will be marked as INJ. Apart from the interjections, the affirmatives such as Hindi 'HAz'('yes') will also be tagged as INJ. Since, this is the only example of such a word, it has been clubbed under Interjections.

h38. ***arre*_INJ**, *tuma A gaye !*
'oh' 'you' 'come' 'have'
“Oh! you have come”

h39. ***hAz*_INJ**, *mEM A gayA*
'yes', 'I' 'come' 'have'
“Yes, I have come”.

5.15 NEG Negative

Negatives like Hindi 'nahIM' (not), 'nA' (no, not), etc. will be marked as NEG. For example,

h40. *vaha Aja* ***nahIM*_NEG** *A pAyegA*
'he' 'today' 'not' 'come' 'will be able'

Also, see examples (b2) and (h25) given above.
Indian languages have reiteration of NEG in certain constructions. For example,

b5. t*umi chobitA dekhbe* ?
'you' 'picture-def' 'will see' ?
“Will you see the picture ?”
b6. *nA*_NEG, *xekhabo nA*_NEG
*'n*o' 'will see (I)' 'not'
“No, I will not see (it)”

The first occurrence of *'nA'* in such constructions will also be marked as NEG.

5.16 UT Quotative

A quotative introduces a quote. Typically, it is a verb. Many Indian languages use quotatives. Given below is an example from Bengali,

b7. *she Ashbe* ***bole*** *bolechilo*
'he' 'will come' 'quotative' 'told'
“He told that he will come”.

5.17 SYM Special Symbol

All those words which cannot be classified in any of the other tags will be tagged as SYM. This tag is similar to the Penn 'SYM'. Also special symbols like $, %, etc are treated as SYM. Since the frequency of occurrence of such symbols is very less in Indian languages, no separate tag is used for such symbols.

5.18 *C Compounds (Make it XC – where X is a variable of the type of the compound of which the current word is a member of)

The issue of including a tag for marking compounds was discussed extensively. Results of algorithms using IIIT-H tag set which included NNC (part of compound nouns) and NNPC (part of proper nouns) showed that these two tags contributed substantially to the low accuracy of the tagger. Since most elements which occur as NNC or NNPC can also occur as NN and NNP, it affected the learning by the machine. So, the question was, why to include tags which contributed more to the errors ? The other aspect, however, was that while human annotators are annotating the data, they know from the context when a certain element is NNC or NN, NNPC or NNP and if marked, this information can be useful for certain applications. The argument is same as the one in favor of including a tag for proper nouns.

Another point which was discussed was that any word class can have compound forms in Indian languages (including adjectives and adverbs).

Therefore, if we decide to have a tag for showing compounds of each type, the number of tags will go very high. The final decision on this was to include a *C tag which will be realised as **catC** tag of the type of compound that the element is a part of. For example, if a certain word is part of a compound noun, it will be marked as NNC, if it is part of a compound adjective, it will be marked as JJC and so on and so forth.
Some examples are given below :

Hindi compound noun *keMdra sarakAra* (Central government) will be tagged as *keMdra_***NNC** *sarakAra_***NN**.

In this example, '*keMdra*' and '*sarakAra*' are both nouns which are forming a compound noun. All words except the last one, of a compound words will be marked as NNC. Thus any NNC will be always followed by another NNC or an NN. This strategy helps identify these words as one unit although they are not conjoined by a hyphen. Similarly, a compound proper noun will be marked as NNPC excluding the last one. eg. *aTala_*NNPC *bihArI_*NNPC *vAjapeyI_*NNP
The first two words, in the above example, will be tagged as NNPC and the last one will be tagged as NNP. Similar to the NNC tag for common nouns, NNPC tag helps in marking parts of a proper noun.

h41. *rAma, mohana aur shyAma ghara gaye.*
'Ram', 'Mohan' 'and' 'Shyam' 'home' 'went'
"Ram, Mohan and Shama went home".

h42. *bagIce meM* ***ranga_*****JJC** ***biraMge_*****JJ** *phUla khile the*
'garden' 'in' 'colourful' 'flowers' 'flowered' 'were'
"The garden had colorful flowers"

Titles such as **Dr., Col., Lt**. etc. which may occur before a proper noun will be tagged as **NNC**. All such titles will always be followed by a Proper Noun. In order to indicate that these are parts of proper nouns but are nonetheless nouns themselves, they will be tagged as NNC, eg. **Col._NNC** Ranjit**_NNPC** Deshmukh**_NNP**

5.19 RDP Reduplication

In this phenomenon of Indian languages, the same word is written twice for various purposes such as indicating emphasis, deriving a category from another category etc. eg. *choTe choTe* ('small' 'small'; very small), *lAla lAla* ('red' 'red'; red), *jaldI jaldI* ('quickyl' 'quickly' ; very quickly)
There are two ways in which such word sequences may be written. They can be written – (a) separated by a space or (b) separated by a hyphen.

The question to be resolved is that in case, they are written as two words (separated by space)– how should they be tagged? Earlier decision was to use the same tag for both the words. However, in this approach, the morphological character of reduplication is missed out. That is, the reduplicated item will then be treated exactly like two independent words of the same category. For example,

h43. *vaha* ***mahaMgI_*****JJ** ***mahaMgI_*****JJ** *cIjZeM kharIda lAyA*
'he' 'expensive' 'expensive' 'things' 'buy' 'bring'
“He bought **all expensive** things”.
h44. *una* ***catura_JJ buddhimAna_JJ*** *baccoM ne samasyA sulajhA lI*
'those' 'smart' 'intelligent' 'children' 'erg' 'problem' 'solved'
“Those **smart** and **intelligent** children solved the problem.

Both (h43) and (h44) have a sequence of adjectives - *mahaMgI_*JJ *mahaMgI_*JJ and *catura_JJ buddhimAna_JJ* respectively. In the first case, the sequence of two adjectives is a case of reduplication (same adjective is repeated twice to indicate the intensity of 'expensive') whereas in the second case the two adjectives refer to two different properties attributed to the following noun. Since reduplication is a highly productive process in Indian languages, it is proposed to include a new tag **RDP** for annotating reduplicatives. The first word in a reduplicative construction will be tagged by its respective lexical category and the second word will be tagged as RDP to indicate that it is a case of reduplication distinguishing it from a normal sequence such as in (h44) above. Some more examples are given underneath to make it more explicit,

h45. *vaha* ***dhIre_*****RB** ***dhIre_*****RDP** *cala rahA thA.*
'he' 'slowly' 'slowly' 'walk' 'PROG' 'was'
“He was walking (very) slowly”.
h46. *usake bAla* ***choTe_*****JJ** ***choTe_*****RDP** *the.*
'his' 'hair' 'short' 'short' 'were'
“He had (very) short hair”
h47. yaha bAta ***galI_*****NN** ***galI_*****RDP** *meM phEla gayI.*
'this' 'talk' 'lane' 'lane' 'in' 'spread' 'went'
“The word was spread in every lane”.

5.20 ECH Echo words

Indian languages have a highly productive usage of echo words such as Hindi '*cAya-vAya*' ('tea' 'echo'), where *'cAya*' is a regular lexical item of Hindi vocabulary and *'vAya*' is an echo word indicating the sense “etc” . These words, on their own, are 'nonsense' words and do not find a place in any dictionary. Thus, the gloss for '*cAya-vAya*' would be '*tea etc*'. It is proposed to add the tag **ECH** for such words.

5.21 UNK Unknown

A special tag to indicate unknown words is also included in the tag set. The annotators can use this tag to mark the words whose category they are not aware of. This tag has to be used very cautiously and sparsely, i.e., only if it is absolutely necessary.

6. Some Special Cases

This section gives the details of certtain aspects of Indian languages which need to be dealt with separately in the tagger. These are issues that cannot be handled by just changing or adding tags.

6.1 'vAlA' type constructions

'*vAlA*' is a kind of suffix used in Hindi and some other Indian languages. It conjoins with nouns (Case I, below) or verbs (Case II) to form adjectives or even nouns. It is also used as an aspectual TAM in a verbal construction (Case III).

h48. *lAla* ***kamIjZa vAlA*** *AdamI merA bhAI hE* .
'red' 'shirt' 'in' 'man' 'my' 'brother' 'is'
“The man in red shirt is my brother”.

h49. *mehanawa* ***karane vAle*** *vyakti ko inAma milegA* .
'hard work' 'doing' 'adj' 'person' 'to' 'prize' 'will get'
The person who works hard will get a prize.

These cases are elaborated below.

Case I: The suffix 'vAlA' can occurr with a noun. For example, ***lAThI vAlA (*** 'stick' 'with' **-**The one with a stick).

h50. *lAThI vAle AdamI ko bulAo*
'with stick' 'man' 'acc' 'call'
“Call the man with the stick”.

This suffix '*vAlA*' in Hindi (a) may be written separately or (b) may be attached to the preceding noun.

(a) In case it is written separately as in '*lAThI vAlA*' above, the word '*lAThI*' will be tagged as NN and the word '*vAlA*' will be tagged as PSP.

The whole expression '*lAThI vAlA*' is an adjective, in which '*lAThI*' is a noun and '*vAlA*' is a suffix which derives an adjective from a noun. Since '*lAThI*' and '*vAlA*' written separately in the above example, they have to be tagged individually. '*vAlA*' in such cases will be treated like a postposition and will be tagges as PSP.

(b) The second possibility is of '*lAThi*' and '*vAlA*' written together as '*lAThIvAlA*'. In such cases it will be treated as one word and will be marked as JJ since 'l*AThIvAlA*' is an adjective.

Case II: '*vAlA*' can also occur after a verb. Example, ***karane vAlA*** ('doing' 'one' – The one who does something)

h51. *mehanata* ***karanevAle ko*** *phala milatA hE*
'hard' 'working one' acc 'fruit' 'get' 'PRES'
The one who works hard gets the fruits".

As mentioned earlier, the suffix '*vAlA*' also joins a verb in its nominal form and makes it an adjective. In this case also, the two words may be written separately (*karane vAle*) or together (*karanevAlA*). In the former case, the two words will be marked as VM and PSP respectively (*karane*_VM *vAle*_PSP). In the latter case, being a single word (*karanevAlA*) it will be tagged as VM (*karanevAle*_VM). It is crucial to retain the 'verb' information in these case, so that at a later stage if we want to annotate its argument structure we should be able to do so (discussed earlier in the document).

Case III: 'vAlA' can also occur as part of TAM. For example,

h52. *mEM wo* ***jAne vAlA*** *hI thA.*
'I' 'particle' 'to go' 'about' 'part' 'was'
"I was about to go"

Although the word '*jAne*' has a '*vAlA*' suffix in (h52) above, the entire expression is not an adjective but is a verb having the aspectual information of 'shortly'. In this case, the sequence '*jAne vAlA*' will be marked as *jAne*_VM *vAlA*_VAUX. The alternative writing convention of writing the sequence as one word (*jAnevAlA*) is possible in this case also. Like the earlier cases, the word will be marked for the category of the content morpheme – which is verb in this case. Thus *jAnevAle* will be tagged as *jAnevAle*_VM.

Here again we stand by our policy that the tag will be decided on the basis of the part of speech and not on the basis of the category of the word in the given sentence(syntactic function). This avoids confusion at the level of manual tagging and aids machine learning as well. So the tag (VM) remain same although the function of the words is different in two different places, it is adjective in Cases I and II and verbal in Case III.

6.2 Honorifics in Indian languages

Hindi (and some other Indian languages) has particles such as '*jI*' or '*sAHaba*' etc. after proper nouns or personal pronouns. These particles are added to denote respect to the referred person. Such honorific words will be treated like particles and will be tagged **RP** like other particles.

h53. ***mantrI*_NN *jI*_RP** *sabhA meM dera se pahuMce .*

'minister'   'hon' 'meeting' 'in' 'late' 'part' 'reached'
“The minister reached late for the meeting”.

6.3 Foreign words

Presence of loan words is a fairly common phenomenon in languages. Most Indian languages have a number of loan word from English. One may also come across words from other Indian languages or Sanskrit in a given text. Such foreign words will be tagged as per the syntactic function of the word in the given context. In special cases, such as when the annotator is not sure of the category of a word, it will be tagged as **UNK.**

7. A Special Note

There may be situations, when an annotator does not feel very confident about the tag for a particular word. The annotator may then assign it different tags in different places. Inconsistency in the manual tagging can affect the learning considerably. Since this is a task which involves a number of human annotators, the methods have to be evolved to check and cross validate the human annotation. Another practical problem in annotation is that in the initial stages of annotation, the annotators need time to get familiar with the tagging scheme and the concept behind each tag. Thus they take some time before coming to a stable stage of decision making for various instances, particularly various ambiguous cases. Especially, in the initial stages, the annotators may often come across cases where their confidence level may not be very high. They may feel the need of some clarifications for these cases. Since the task of annotation has to go on and immediate clarification may not be possible, the annotators may be forced to take decisions and mark a case as they consider appropriate at that point of time. Over a period of time, with better understanding of the tags and tagging scheme, they may reach a stable stage. However, by then they may already have tagged a given case differently in different places thus introducing inconsistency in the annotated corpus. At a later stage, it will be difficult to go back to all the cases that have been annotated by then and correct them. So the chances are that the annotators may proceed with the revised decision and leave the earlier annotation as such. This will introduce inconsistencies in the annotated corpus.

To control such a situation, it is decided to provide a way by which the annotators can initially mark the uncertainty of their decision so that they can easily extract these cases easily and take them up for discussions and clarifications.

This 'uncertainty' will be annotated as follows :

a) The annotators first mark such a case with a tag that they consider appropriate at the time of annotation.

b) Along with the chosen tag, they also put a question mark (?) against that tag. The question mark will indicate that this case is not yet resolved and will be finalized after clarification or discussion.
c) All the cases with a question mark can be later taken out and placed for discussion. An annotator will be responsible for bringing such cases for discussion and once the cases are resolved, the annotator will go back and correct the tag. In case the tag assigned by the annotator initially itself is correct, the annotator will remove the question mark against it.

> **This is a purely temporary measure and the data finally submitted by an annotator should not have any words having a question mark.**

8. Chunk Tags Chosen for the Current Scheme

This section deals with the chunk tags. Not many of the issues discussed above hold for defining the chunk tags. Various points which have been deliberated upon in relation to chunking scheme are :

1. Definition of a chunk
2. Chunk Types
3. Some Special Cases
4. Annotation method/procedure

8.1 Definition of a chunk

Following issues related to the definition of a chunk were discussed :
What constitutes a 'chunk' ?

A typical chunk consists of a single content word surrounded by a constellation of function words (Abney,1991). Chunks are normally taken to be a 'correlated group of words'.

The next issue, however, is - How to define the boundaries of these 'correlated word groups' for our purpose?

For example, which case in the following pairs should be grouped as a chunk ?

((*xillI meM*)) OR ((*xillI*)) *meM*
'Delhi' 'in' *'Delhi' 'in'*
((*rAjA kA betA*)) OR ((*rAjA kA*)) ((*betA*))
'king' 'of' 'son' 'king 'of' 'son'
((*rAjA ke bete kI paxnI*)) OR ((*rAjA ke*)) ((*bete k*I)) ((*paxnI*))
'king' 'of' 'son' 'of' 'wife' 'king' 'of' 'son' 'of' 'wife'

Following definition of a 'chunk' was evolved through discussion :

“A minimal (non recursive) phrase(partial structure) consisting of correlated, inseparable words/entities, such that the intra-chunk dependencies are not distorted”. Each chunk type discussed and the decided upon is described below .

8.2. Chunk Types

Based on the above definition of chunk, issues related to various chunk types were discussed. A chunk would contain a 'head' and its modifiers.

8.2.1 NP Noun Chunk

Noun Chunks will be given the tag NP and include non-recursive noun phrases and postpositional phrases. The head of a noun chunk would be a noun. Specifiers will form the left side boundary for a noun chunk and the vibhakti or head noun will mark the right hand boundary for it. Descriptive adjective/s modifying the noun will be part of the noun chunk. The particle which anchors to the head noun in a noun chunk will also be grouped within the chunk. If it occurs after the noun or vibhakti, it will make the right boundary of the chunk. Some example noun chunks are :

((*bacce*_NN))_NP, ((*kucha*_QF *bacce*_NN))_NP,
'children' 'some' 'children'
((*kucha*_QF *acche*_JJ *bacce*_NN))_NP, ((*Dibbe*_NN *meM*_PSP))_NP,
'some' 'good' 'children' 'box' 'in'
((eka_QC *kAlA*__JJ *ghoDZA*_NN))_NP ,
'one' 'black' 'horse'
((*yaha*_DEM *nayI*_JJ *kitAba*_NN))_NP,
'this' 'new' 'book'
((**isa**_DEM *nayI*_JJ *kitAba*_NN *meM*_PREP))_NP,
'this' 'new' 'book' 'in'
((*isa*_DEM *nayI*_JJ *kitAba*_NN *meM*_PSP *bhI*_RP))_NP
'this' 'new' 'book' 'in' 'also'

The issue of genitive marker and its grouping with the nouns that it relates to was discussed in detail. For example, the noun phrase '*rAma kA beTA*' contains two nouns '*rAma*' and '*beTA*'. The two nouns are related to each other by the vibhakti '*kA*'. The issue is whether to chunk the two nouns separately or together? Linguistically, '*beTA*' is the head of the phrase “*rAma kA beTA*”. '*rAma*' is related to '*beTA*' by a genitive relation which is expressed through the vibhakti '*kA*'. Going by our definition of a 'chunk' we should break '*rAma kA beTA*' into two chunks (((*rAma kA*))_NP, ((*beTA*))_NP) by breaking '*rAma kA*' at 'kA' vibhakti . Moreover, if we chunk 'rAma kA beTA' as one chunk, linguistically, we will end up with a recursive noun phrase as a single chunk ((((*rAma kA)) beTA*)) which also is against our definition of a chunk.

Therefore, it was decided that the genetive markers will be chunked along with the preceding noun. Thus, the noun group 'rAma kA beTA' would be chunked into two chunks.

h54. **((*rAma kA*))NP** ((*beTA*))NP acchA hE “Ram's son is good”
h55. ((*kitAba*))NP **((*rAma kI*))NP** hE “The book belongs to Ram”

For the noun groups such as “*usakA beTA*” it was decided that they should be chunked together.

8.2.2 Verb Chunks

The verb chunks would be of several types. A verb group will include the main verb and its auxiliaries, if any. Following are some examples of verb chunks from Hindi,

((*khAyA*)), ((*khA rahA hE*)), ((*khA sakawe hEM*))
'ate' 'eat' 'PROG' 'is' 'eat' 'can' 'PRES'

The types of verb chunks and their tags are described below.

8.2.2.1 VGF Finite Verb Chunk

As mentioned in 5.4 above, a verb group sequence (V VAUX VAUX . .) contains a main verb and its auxiliaries. The group itself can be finite or non-finite. In case of it being finite, the main verb in such a sequence may not be finite. The finiteness is known by the auxiliaries. Therefore, it is decided to mark the finiteness of the verb at the chunk level. Thus, any verb group which is finite will be tagged as **VGF.** For example,

h56. *mEMne ghara para khAnA ((khAyA*_VM*))*_**VGF**
'I erg' 'home' 'at' 'meal' 'ate'

h57. *vaha cAvala ((khA*_VM *rahA*_VAUX *hE*_VAUX*))*_**VGF**
'he' 'rice' 'eat' 'PROG' 'is'

8.2.2.2 VGNF Non-finite Verb Chunk

A non-finite verb chunk will be tagged as **VGNF.** For example,

h15a) *seba* ***((khAtA*_VM *huA*_VAUX))_VGNF** *laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

h16a) *mEMne* ***((khAte – khAte*_VM*))*_VGNF** *ghode ko dekhA*
'I erg' 'while eating' 'horse' acc 'saw'
h17a) *mEMne ghAsa* ***((khAte*_VM *hue*_VAUX))_VGNF** *ghoDe ko dekhA*
'I erg' 'grass' 'eating' 'PROG' 'horse' acc 'saw'

The IIIT-H tagset had initially included three tags for the non-finite verbal forms. Unlike Penn tagset, all non finite verbs, which are used as adjectives, were marked as VJJ at the POS level. Similarly, to mark adverbial non-finite verbs, the POS tagset had VRB tag. A tag VNN was included to mark the nominalized verbs.

However, during the discussions IL standards, it was pointed out that inclusion of too many finer tags hampers machine learning. Moreover, the marking is based on syntactic information, which we should avoid at the POS level, unless it is contributing to further processing in a substantial way. On the other hand, it is important to mark finite non-finite distinction in a verbal expression as it is a crucial information and is also easy to learn. As discussed under 5.4 above, it was decided to mark this distinction at the chunk level, rather than at the POS level. Therefore, the tag VGNF has been included to mark non-finite adverbial and adjectival verb chunk.

8.2.2.3 VGINF Infinitival Verb Chunk

This tag is to mark the infinitival verb form. In Hindi, both, gerunds and infinitive forms of the verb end with a *-nA* suffix. Since both behave functionally in a similar manner, the distinction is not very clear. However, languages such as Bangla etc have two different forms for the two types. Examples from Bangla are given below.

b8. *Borabela* ***((snAna karA))_*** **VGNN** *SorIrera pokze BAlo*
'Morning' 'bath' 'do-verbal noun' 'health-gen' 'for' 'good'
'Taking bath in the early morning is good for health"

b9. *bindu Borabela ((* ***snAna karawe))_*** **VGINF** *BAlobAse*
'Bindu' 'morning' 'bath' 'take-inf' 'love-3pr'
"Bindu likes to take bath in the early morning"

In Bangla, the gerund form takes the suffix *–A / -Ano*, while the infinitive marker is *–we*. The syntactic distribution of these two forms of verbs is different. For example, the gerund form is allowed in the context of the word *darakAra* "necessary" while the infinitive form is not, as exemplified below:

b10 *Borabela ((snAna karA))_* **VGNN** *darakAra*
'Morning' 'bath' 'do-verbal noun' 'necessary'
"It is necessary to take bath in the early morning"

b11. **Borabela ((snAna karawe))_* **VGINF** *darakAra*

Based on the above evidence from Bangla, the tag *VGINF* has been included to mark a verb chunk.

8.2.2.4 VGNN Gerunds

A verb chunk having a gerund will be annotated as **VGNN**. For example,

h18a. *sharAba* *((***pInA***_**VM))_VGNN** *sehata ke liye hAnikAraka hE.*
'liquor' 'drinking' 'heath' 'for' 'harmful' 'is'
“Drinking (liquor) is bad for health”

h19a. *mujhe rAta meM ((***khAnA***_**VM))_VGNN** *acchA lagatA hai*
'to me' 'night' 'in' 'eating' 'good' 'appeals'
“I like eating at night”

h20a. ((***sunane***_**VM** ***meM***_**PSP***))*_**VGNN** *saba kuccha acchA lagatA hE*
'listening' 'in' 'all' 'things' 'good' 'appeal' 'is'

8.2.3 JJP Adjectival Chunk

An adjectival chunk will be tagged as **JJP**. This chunk will consist of all adjectival chunks including the predicative adjectives. However, adjectives appearing before a noun will be grouped together with the noun chunk. A JJP will consist of phrases like

h58. *vaha laDaZkI hE*((*suMdara*_**JJ** *sI*_RP))_**JJP**
'she' 'girl' 'is' 'beautiful' 'kind of'

h59. *hAthI AyA* ((*moTA*_*C *tagadA*_**JJ**))_**JJP**
'elephant' 'came' 'fat' 'powerful'

h60. *vaha laDakI ((bahuta*_INTF *sundara*_**JJ**))_**JJP** *hE*
'she' 'girl' 'very' 'beautiful' 'is'

Cases such as (h61) below will not have a separate JJP chunk. In such cases, the adjectives will be grouped together with the noun they modify. Thus forming a NP chunk.

h61. ((*kAle*_**JJ** *ghane*_**JJ** *laMbe*_**JJ** *bAla*_**NN**))_**NP**
'black' 'thick' 'long' 'hair'

8.2.3.1 Some special cases

Following examples from Hindi present a

h62. *xillI meM* ***rahanevAlA*** *merA BAI kala A rahA hE .*
'Delhi' 'in' 'staying' 'my' 'brother' 'tomorrwo' 'come' 'PROG' 'is'
“My brother who stays in Delhi is coming tomorrow”.
h63. *usane Tebala para* ***rakhA huA*** *seba khAyA.*
'(s)he erg' 'table' 'on' 'kept' 'apple' 'ate'

"He ate the apple kept on the table".

In (h62) above '*rahanevAlA*' is an adjectival participle. But we do NOT mark it as JJP. Instead, it will be marked as a **VGNF**. The decision to tag it as a VGNF is based on the fact that such adjectival participles are derived from a verb can have their arguments. This information is useful for processing at the syntactic level. Thus, '*rahanevAlA*' in (h62) will be annotated as follows:

h62a. *xillI meM* ((***rahanevAlA*_VM*)*_VGNF** *merA BAI kala A rahA hE* .

Similarly, in (h63) above, the chunk '*rakhA huA*' is an adjective but will also be marked as a VGNF since this also derived from a verb and chunks like '*Tebala pra*' etc are its arguments. So the chunk name will be **VGNF** and the POS tag will be **VM** which might be followed by an auxiliary verb tagged as **VAUX**. (h63a) shows how '*rakhA huA*' will be annotated :

h63a. *usane Tebala para ((*___rakhA___**_VM** ***huA*_VAUX*))*_VGNF***seba khAyA*.

8.2.4 RBP Adverb Chunk

This chunk name is again in accordance with the tags used for POS tagging. This chunk will include all pure adverbial phrases.

h64. *vaha* ((*dhIre-dhIre*_**RB**))**_RBP** *cala rahA thA.*
'he' 'slwoly' 'walk' 'PROG' 'was'
"He was walking slowly"

Now consider the following examples:

h65. *vaha* ***dagamagAte hue*** *cala rahA thA* .
'he' ' 'walk' 'PROG' 'was'
"He was walking

h66. *vaha khAnA* ***khAkara*** *ghara gayA* .
'he' 'meal' 'after eating' 'home' 'went'
"He went home after eating his meal"

In the above examples, '*dagamagAte hue*' and 'k*hAkara*' are non finite forms of verbs used as adverbs. Similar to adjectival participles these will also be chunked as **VGNF** and not as **RBP**. The reason for this is that we need to preserve the information that these are underlying verbs. This will be a crucial information at the level of dependency marking where the arguments of these verbs will also be marked.

((isa_PRP nayI_JJ kitAba_NN meM_PSP bhI_RP))_NP
'this' 'new' 'book' 'in' 'also'

8.2.5 NEGP Negatives

(i) In case a negative particle occurs around a verb, it is to be grouped within verb group. For example,

h67. *mEM kala dillI ((**nahIM**_*NEG ***jA**_VM ***rahI**_VAUX*))_*VGF
"I" "tomorrow" "Delhi" "not" "go" "Cont"

h68. ((***binA***_NEG ***bole***_VM))_VGNF *kAma ((**nahIM**_*NEG ***calatA**_VM*))_*VGF
"without" "saying" "work" "not" "happen"

However ,

h69. **binA** kucha **bole** kAma **nahIM calatA**
"without" "something" "saying" "work" "not" "happen"

In the above sentence, the noun "*kucha*" is coming between the negative "*binA*' and verb "*bole*". Here, it is not possible to group the negative and the verb as one chunk. At the same time, "*binA*" cannot be grouped within an NP chunk, as functionally, it is negating the verb and not the noun. To handle such cases an additional **NEGP** chunk is introduced. If a negative occurs away from the verb chunk, the negative will be chunked by itself and chunk will be tagged as NEGP. Thus,

h69a. **((*binA*))_NEGP** **((**(*kucha*))_NP **((*bole*))_VG** **((**(*kAma*))_NP ((*nahIM calatA*))_VG

8.2.6 CCP Conjuncts

Conjuncts are functional units information about which is required to build the larger structures. Take the following examples of cunjunct usages :

h70. (*rAma kitAba paDha rahA thA)* ***Ora*** *(mohana Tennisa khela rahA thA).*
"Ram was reading a book **and** Mohan was playing tennis"

h71. (*rAma ne batAyA)* ***ki*** *(usakI kitAba acchI hE).*
"Ram said **that** his book is good"

h72. (*rAma)* **Ora** *(mohana) Tennisa khela rahe the.*
"Ram **and** Mohan were playing tennis".

h73. (*merA bhAI rAma)* ***Ora*** *(usakA dosta mohana) Tennisa khela rahe the.*
"My brother Ram **and** his friend Mohan were playing tennis".

h74 . *rAma (saphZeda kapade)* ***Ora*** *(nIle jute) pahane thA.*
"Ram was wearing white clothes and blue shoes".

h75. *rAma eka (halkI)* ***Ora*** *(nIlI) bOla lAyA.*
"ram brought a light **and** blue ball".

The sentences above have various types of conjoined structures. To represent these conjoined structures, it is decided to form separate chunks for conjuncts and the elements a conjunct conjoins. Thus (h70) and (h71) above will be chunked as (h70a) and (h71a) given below,

h70a. ((*rAma*))_NP ((*kitAba*))_NP ((*paDha rahA thA*))_VG **((*Ora*))CCP** ((*mohana*))_NP ((*Tennisa*))_NP ((*khela rahA thA*))_VG.

h71a. ((*rAma ne*))_NP ((*batAyA*))_NP **((*ki*))_CCP** ((*usakI*))_NP ((*kitAba*))_NP ((*acchI*))_JJP ((*hE*))_VG.

Expression '*rAma Ora mohana*' in example (h72) is a complex NP. Though complex, the expression can be annotated as a single NP chunk as functionally it is the subject of the verb 'play'. However, example (h73) presents a case where it would be better to form three independent chunks for the complex subject NP. Though the conjunct '***Ora***' is conjoining '*rAma*' and '*mohana*', both '*rAma*' and '*mohana*' have their respective modifiers. To make it explicit, it is better to treat them as two independent NP chunks conjoined by a CCP.

h73a. ((*merA bhAI rAma*)) **((*aura*))_CCP** ((*usakA dosta mohana*))_NP ((*Tennisa*))_NP ((*khela rahe the*))_VG.

Following this, the subject NP of (h72) would also be annotated similarly. Therefore,

h72a. ((*rAma*))_NP **((*aura*))_CCP** ((*mohana*))_NP ((*Tennisa)*)_NP ((*khela rahe the*))_VG.

The annotation for cases such as (h74) and (h75) would be as follows :

h74a. ((*rAma*))_NP ((*safeda kapade*))_NP **((*aura*))_CCP** ((*nIle jute*))_NP ((*pahane thA*))_VG.

h75a. ((*rAma*))_NP ((*eka*))_JJP ((*halkI*))_JJP **((*aura*))_CCP** ((*nIlI*))_JJP ((*bOla*))_NP ((*lAyA*))_VG

Thus the decision for conjuncts is - the conjoined entities will be broken into separate chunks. eg. ((*rAma*))_NP ***((Ora))*_CCP** ((*SyAma*))_NP

8.2.7 FRAGP Chunk Fragments

Some times certain fragments of chunks are separated from the chunks to which they belong. For example :

h76. ***rAma*** (*jo merA baDZA bhAI hE*) ***ne*** *kahA ...*
'Ram' 'who' 'my' 'elder' 'brother' 'is' 'erg' 'said'

In the above example, vibhakti '*ne*', which is a case marker of the noun '*rAma*', is separated from it by an intervening clause. Syntactically, '*ne*' is a part of the noun chunk '*rAma ne*'. However, at times it can be written separately. The following was decided for such fragments :

(i) There will be a separate chunk for the vibhakti in constructions where it gets separated from the noun it would normally be grouped with. This chunk can have more than one entity within it.

h77. ((***rAma***))**_NP,** *mere dillI vAle bhAI,* **((*ne*))_FRAGP** *kahA*

'Ram' 'my' 'Delhi' 'from' 'brother' 'erg' 'said'

(ii) If the entities embedded between the noun and it's vibhakti are a series of nouns the entire group will be chunked as a single noun chunk.

h78. **((***isa* ***'upanyAsa samrATa'*** *Sabda kA***))_NP**

'this' 'Novel' 'King' 'word' 'of'

8.2.8 BLK Miscellaneous entities

Entities such as interjections and discourse markers that cannot fall into any of the above mentioned chunks will be kept within a separate chunk.

eg. ((*oh*_INJ))**_BLK,** ((*arre*_INJ))**_BLK**

8.3 Some Special Cases

Apart from the above, some special cases related to certain lexical types are discussed below.

8.3.1 Conjunct Verbs

The issue whether to treat the noun/adjective which is part of a conjunct verb differently by marking it with a special tag (NVB/JVB) or to treat it as a noun like any other noun at the POS level was deliberated on.
The question was based on the following observations :

a) NVB/JVB , as part of conjunct verbs, are most often not recognized by the learning algorithms.

b) Having NVB at the POS level is based on syntactic considerations. Therefore, do we really need to go for it ? Instead, at the POS level we mark the noun as a noun and leave the decision of marking a conjunct verb as single unit for a later level.

c) Moreover, since the noun, which is part of a conjunct verb (Kriyamula), can occur away from its 'verbaliser', it becomes difficult to differentiate it from a 'noun' which may be an argument of the verb. This also creates problem for chunking of the verb group. The two components of the chunk have to be separately marked and have to be joined at the syntactic level.

d) If NVB is marked at the POS level, a natural consequence would be to group it with its verbaliser as a VG chunk. In fact, that is the purpose of identifying it as different from a noun. However, sometimes one comes across expressions such as '*mEMne unase eka* ***prashna kiyA***' (I posed a question to him). In this sentence, '***eka***' is a modifier of '*prashna*'. '*prashna karana*' is recognized as a conjunct verb in Hindi by most Hindi speakers. Following example shows the problem of grouping '*praSna karanA*' as a single VG:

POS : *mEMne*_PRP *unase*_PRP *eka*_QC *prashna*_NVB *kiyA*_VM
Chunk : ((*meMne*_PRP))_NP ((*unase*_PRP))_NP ((*eka*_QC))_JJP ((*prashna*_NVB *kiyA*_VM))_VGF

Once "*praSna karanA*" are grouped together as a chunk, it will be difficult to show the relation between '*eka*' and 'prashna' subsequently.

Thus, an alternative was proposed wherein, the noun of the conjunct verb is tagged as NN at the POS level which is accordance with the decision to tag the lexical item based on its lexical category. Thereafter, the noun is grouped with its preceding adjectival modifiers as an NP chunk. The only problem in this approach is that the information of a noun verb sequence being a conjunct verb is not captured till the chunk level and the noun of the conjunct verb is separated from its verbaliser. However, the approach has following advantages :

1) At the POS level, the word is tagged for its grammatical category and not for its syntactic function. This eases the decision making at the POS level. And marking the information, that the conjunct verbs which are composed of two words form one lexeme semantically, is postponed to a later level.

2) It allows us to show the modifier-modified relation between an adjective such as '*eka*' in the above example with its modified noun '*praSna*'.

3) Since the information of a noun verb sequence being a 'kriyamula' is crucial at the syntactic level, it will be captured at that level by marking the relation between the 'noun' and its verbaliser by an appropriate tag. Therefore, the decision is :

The noun/adjective and verb (internal components of a conjunct verb) will be chunked separately.
eg. *prashna karanA* **- ((*prashna*))NP ((*kiyA*))VG**
ucita kiyA - **((*ucita*))JJP ((*kiyA*))VG**

8.3.2 Particles

Regarding the particles, it was decided that the particles will be chunked with the same chunk as the anchor word they occur with. Thus,
eg. ((*rAma ne* ***bhI***))**_NP, ((*mEM* *wo*))_NP,**

'Ram' 'erg' 'also' 'I' 'emph'

8.3.3 Quantifiers

The issue of chunking quantifiers was discussed in great details. Numbers can occur (a) as noun modifiers before a noun (***haZaroM*** *ladakoM ne – 'thousands' 'boys' 'erg'*) or (b) can occur without a noun (***hazAroM*** *ne – 'thousands' 'erg'*) with a nominal inflection. The issue of whether to treat the quantifiers of the type (b) as nouns was discussed. The issue is whether (b) is a case of an ellipsis of the noun after a number or whether it is the number itself which is the noun. If the latter has to be followed then the POS tag for quantifiers in such cases should be NN. Following decisions were taken :

(i) A 'QC' or 'QO' occuring with a noun will be part of the noun chunk.

h79. ((***hazAroM*****_QC** *logoM_NN* *ne*_PSP))_NP *yaha driSya dekhA*
'thousands' 'people' 'erg' 'this' 'scene' 'watched'
“Thousands of people watched this scene”.

h80. ((***dUsare*****_QO** *ladake*_NN *ne*_PSP))_NP *isa samasyA ko sulajhA diyA*
'second' 'boy' 'erg' 'this' 'problem' acc 'solve' 'did'
“The second boy solved this problem”.

(ii) All categories occurring without a noun, with nominal inflections (overt or otherwise) will be tagged as noun.

h81. ((***hazAroM_NN*** *ne*_PSP))_NP *yaha driSya dekhA*
'thousands' 'erg' 'this' 'scene' 'watched'
“Thousands watched this scene.

h82. **((*mote*_NN *ne*_PSP))_NP ((chote_NN ko_PSP))_NP ((***mArA***))_**VGF
'fat' 'erg' 'small' 'to' 'killed'

8.3.4 Punctuations

All punctuations, with an exception of sentence boundary markers and clausal conjuncts, will be included in the preceding chunk. For example

h83. ((*usane*_PRP))_NP ((***kahA*****_VM –_SYM)_**VGF
'He erg' 'said'
((***"*****_SYM** ***yaha*****_PRP**))_NP ((*Thika*_JJ))_JJP ((***hE*****_VM "_SYM**))_VGF
'this' 'proper' 'is'
“He said, “this is not right” “.

h84. *rAma AyA* **((,_SYM))_CCP** *mohana gayA*
'Ram' 'came' , 'Mohan' 'went'
“Ram came and Mohan left”.

Punctuations such as (a) hyphens and (b) quote marks will be taken care of by the tokenizer.

(a) Hyphens: Identified to be of two types:-

- Without space on either sides, as in the case of compound nouns
  eg. *mAtA-pitA(mother-father)*

- With spaces, as in the case of

  h85. *rAma ne kahA – yaha thIk hE*
  'Ram' 'erg 'said' – 'this' 'proper' 'is'

(b) Quote marks (single and double both) : Identified to be of two types:-
(i) opening
(ii) closing

9. Annotation Procedure

To maintain consistency in the data format and the annotation, it was decided to use 'Sanchay', a facility developed at IIIT, Hyderabad for the annotation task.

10. Conclusion

The annotation standards for POS tagging and chunking for Indian languages include 26 tags for POS (Table-1 in Appendix) and 11 chunk tags (Table-2 in Appendix. The tags are decided on coarse linguistic information with an idea to expand it to finer knowledge if required.

11. References

Steven Abney. Parsing by Chunks. In: Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing.* Kluwer Academic Publishers, Dordrecht. 1991.

Following participated in the meetings/discussions :

IIIT, Hyderabad – Rajeev Sangal, Dipti M Sharma, Soma Paul, Lakshmi Bai, Research students of LTRC, IIIT, Hyderabad.
University of Hyderabad, Hyderabad – Amba Kulkarni, G. Uma Maheshwar Rao, Rahmat Yousufzai
IIT, Bombay – Pushpak Bhattacharya, Om Damani, Rajat Kumar Mohanty, Sushant S Develkar, Pranesh, Maneesh
IIT, Khadagapur – Sudeshna Sarkar, Anupan Basu, Pratyush Banerjee, Sandipan Dandapat
CDAC, Pune – Saurabh Singhal, Abhishek Gupta, Ritu Bara, Mahendra Pandey
CDAC, Noida – Vijay Kumar, K K Arora
IIIT, Allahabad – Ratna Sanyal
Tamil University, – S. Rajendran
AUKBC, Chennai – Sobha Nair

Jadhavepur University – Shivaji Bandopadhyay

12. Acknowledgements

Ms Pranjali Karade prepared the initial document describing IIIT-H tagging scheme which has been an immense help in preparing the current document. Thanks to Soma Paul and Vasudhara Sarkar for providing Bangla examples given in the text.

13. Appendix

13.1. POS Tag Set for Indian Languages (Nov 2006, IIIT Hyderabad)

Sl No.	**Category**	**Tag name**	**Example**
1.1	Noun	NN	
1.2	NLoc	NST	
2.	Proper Noun	NNP	
3.1	Pronoun	PRP	
3.2	Demonstrative	DEM	
4	Verb-finite	VM	
5	Verb Aux	VAUX	
6	Adjective	JJ	
7	Adverb	RB	*Only manner adverb
8	Post position	PSP	
9	Particles	RP	bhI, to, hI, jI, hA.N, na,
10	Conjuncts	CC	bole (Bangla)
11	Question Words	WQ	
12.1	Quantifiers	QF	bahut, tho.DA, kam (Hindi)
12.2	Cardinal	QC	
12.3	Ordinal	QO	
12.4	Classifier	CL	
13	Intensifier	INTF	
14	Interjection	INJ	
15	Negation	NEG	
16	Quotative	UT	ani (Telugu), endru (Tamil), bole/mAne (Bangla), mhaNaje (Marathi), mAne (Hindi)
17	Sym	SYM	
18	Compounds	*C	
19	Reduplicative	RDP	
20	Echo	ECH	
21	Unknown	UNK	

It was decided that for foreign/unknown words that the POS tagger may give a tag ”UNK”

13.2. Chunk Tag Set for Indian Languages

Sl. No	***Chunk Type***	***Tag Name***	***Example***
1	Noun Chunk	NP	*Hindi: ((merA nayA* **ghara))_NP** *"my new house"*
2.1	Finite Verb Chunk	VGF	*Hindi: mEMne ghara para khAnA ((khAyA*_VM*))*_**VGF**
2.2	Non-finite Verb Chunk	VGNF	*Hindi: mEMne* ***((khAte – khAte*_VM*))*_VGNF** *ghode ko dekhA*
2.3	Infinitival Verb Chunk	VGINF	*Bangla : bindu Borabela* ***((snAna karawe))*_VGINF** *BAlobAse*
2.4	Verb Chunk (Gerund)	VGNN	*Hindi: mujhe rAta meM ((***khAnA*_VM))_VGNN** *acchA lagatA hai*
3	Adjectival Chunk	JJP	*Hindi: vaha laDaZkI hE*((*suMdara*_**JJ** *sI*_RP))_**JJP**
4	Adverb Chunk	RBP	*Hindi : vaha* ((*dhIre-dhIre*_**RB**))_**RBP** *cala rahA thA*
5	Chunk for Negatives	NEGP	*Hindi:* **((*binA*))_NEGP** **((***kucha*))_NP **((*bole*))_VG** **((***kAma*))_NP ((*nahIM calatA*)**)_VG**
6	Conjuncts	CCP	Hindi: ((*rAma*))_NP ***((Ora))*_CCP** ((*SyAma*))_NP
7	Chunk Fragments	FRAGP	Hindi; ***rAma*** (*jo merA baDZA bhAI hE*) ***ne*** *kahA...*
8	Miscellaneous	BLK	

Created By : Dipti M Sharma <dipti@iiit.ac.in>
Last Revision By : Dipti Misra Sharma
Creation Date : 30-11-2006
Revision Dates : 15-12-2006

Jawaharlal Nehru University, New Delhi

Bureau of Indian Standards Tagset

Indian Languages Corpora Initiative

DeitY-Sponsored project

The Bureau of Indian Standards (BIS) tagset was standardized in 2010 to enable Part-Of-Speech (POS) annotation of the scheduled languages of the Indian Constitution.

Parts of Speech (PoS) Annotation Guidelines
Following the BIS Tagset
Version 1.1

The tagset below describes the comprehensive tags list for the 12 Indian languages in Phase 1. Some changes were brought to the initial tagging decisions at the ILCI-ILMT POS workshop in July 2012. Those have been documented at the end of the initial document (p. 17)

The Bureau of Indian Standards (BIS) Tagset has recommended the use of a common tagset for the part of speech annotation of Indian languages. The tagset, incorporating the advice of the experts and the stakeholders in the area of natural language processing and language technology of Indian languages, has to be followed in the annotation tasks taking place in Indian languages after August, 2010.

The annotations taking place under the Indian Languages Corpora Initiative (ILCI) program is following the BIS tagset as proliferated.

This document is an attempt to present a guideline for annotation using the BIS tagset. The BIS tagset has a total of 38 annotation level tags which are common to all the Indian languages covered under this tagset.

Sl. No	**Category**			**Label**	**Annotation Convention****	**Remarks**
	Top level	Subtype (level 1)	Subtype (level 2)			
1	**Noun**			**N**	**N**	
1.1		Common		NN	N__NN	
1.2		Proper		NNP	N__NNP	
1.3		Verbal		NNV	N__NNV	The verbal noun ub type is only for languages such as Tamil and Malyalam)
1.4		Nloc		NST	N__NST	
2	**Pronoun**			**PR**	**PR**	
2.1		Personal		PRP	PR__PRP	
2.2		Reflexive		PRF	PR__PRF	
2.3		Relative		PRL	PR__PRL	
2.4		Reciprocal		PRC	PR__PRC	
2.5		Wh-word		PRQ	PR__PRQ	
2.6		Indefinite		PRI	PR__PRI	
3	**Demonstrative**			**DM**	**DM**	
3.1		Deictic		DMD	DM__DMD	

3.2		Relative		DMR	DM__DMR	
3.3		Wh-word		DMQ	DM__DMQ	
		Indefinite		DMI	DM__DMI	
4	**Verb**			**V**	**V**	
4.1		Main		VM	V__VM	
04/01/01			Finite	VF	V__VM__VF	
04/01/02			Non-finite	VNF	V__VM__VNF	
04/01/03			Infinitive	VINF	V__VM__VINF	
04/01/04			Gerund	VNG	V__VM__VNG	
4.2		Auxiliary		VAUX	V__VAUX	
5	**Adjective**			**JJ**		
6	**Adverb**			**RB**		Only manner adverbs
7	**Postposition**			**PSP**		
8	**Conjunction**			**CC**	**CC**	
8.1		Co-ordinator		CCD	CC__CCD	
8.2		Subordinator		CCS	CC__CCS	
8.3		Quotative		UT	CC__CCS__UT	
9	**Particles**			**RP**	**RP**	
9.1		Default		RPD	RP__RPD	
9.2		Classifier		CL	RP__CL	
9.3		Interjection		INJ	RP__INJ	
9.4		Intensifier		INTF	RP__INTF	
9.5		Negation		NEG	RP__NEG	
10	**Quantifiers**			**QT**	**QT**	
10.1		General		QTF	QT__QTF	
10.2		Cardinals		QTC	QT__QTC	
10.3		Ordinals		QTO	QT__QTO	
11	**Residuals**			**RD**	**RD**	
11.1		Foreign word		RDF	RD__RDF	A word written in script other than the script of the original text

11.2		Symbol		SYM	RD__SYM	For symbols such as $, & etc
11.3		Punctuation		PUNC	RD__PUNC	Only for punctuations
11.4		Unknown		UNK	RD__UNK	
11.5		Echowords		ECH	RD__ECH	

** The annotation is to be done using the lowest level tag of the type hierarchy. Once the lower level tag is selected, the higher level tags should be stored automatically.

Table 1 above shows all the eleven categories and their respective sub-categories followed by the part-of-speech labels for the subcategories. Except for the three categories of adjective, adverb and postposition, all the categories have some two or more sub-categories. The category of residual is though not part of the language, it is part of the text which is to be annotated. Therefore, this category also has extra-linguistic elements appearing in the text sub-categorized.

Syntactic Function vs. Lexical Category

A word belonging to a particular lexical category may function differently in a given context. For example, the lexical category of हरिजन in Hindi is a noun. However, functionally, हरिजन is used as an adjective in ex.1 below,

Ex.1. एक दिन पाँच बजे खबर आयी कि कोई हरिजन बालक उनसे मिलना चाहता है
One day five o'clock news came that some harijan boy him meet wants is
One day, a message came at five o'clock that a 'harijana' boy wanted to meet him.

Such cases require a decision on whether to tag a word according to its lexical category or by its syntactic category. Since the word in a context has syntactic relevance, it appears natural to tag it based on its syntactic information. However, such a decision may lead to further complications.

In annotation, the syntactic function of a word is not considered for POS tagging. Since the word is always tagged according to its lexical category there is consistency in tagging. This reduces confusion involved in manual tagging. Also the machine is able to establish a word-tag relation which leads to efficient machine learning.

In short, it is recommended that syntactic and semantic/pragmatic functions were not to be the basis of deciding a POS tag.

1. Noun (N)

The top level category of Noun has been divided into three sub-categories- common, proper and spatio-temporal. Common nouns are pretty easy to decipher, however there has been some discussion on what to include within proper and spatio-temporal nouns.

1.1. Common Nouns (N_NN)

The nouns that simply function as noun and are content words should be marked as the common noun. This includes the general variety of all the nouns, e.g. काम, आम, शहर, जानवर etc.

1.2 Proper Noun (N_NNP)

The proper nouns are basically some specific names which denote to one particular entity. It includes the names of person, place or thing. The examples would be राम, मोहन, कोलकाता, दिल्ली, हिमालय, कोका-कोला etc.

No separate tag has been assigned to abbreviations. They are to be marked as proper nouns. Acronyms, if used as proper nouns, should be marked as proper nouns and if common then as common.

बीजेपी/N_NNP, भाजपा/N_NNP

It was unclear to determine name of class against name of instance, for example: apple, mango, grape as instances under the class of fruit. Under these cases, certain decisions were agreed upon. Names of cities, institutions, organizations, people and months were taken as NNP whereas names of medicines, diseases, flowers, animals and seasons are to be taken as NN.

1.3 Spatio-Temporal Nouns (N_NST)

There is a specific set of words which function sometimes as argument of verb and sometimes as the postposition. The sub-category of NST in the standards document has been included to specifically capture these words. In these words are as follows:
आगे, पीछे, ऊपर, नीचे, बाद, पहले, अंदर and बाहर
So, irrespective of their syntactic function, these words have to be constantly marked as the NST.

वह आगे/NN_NST जा रहा था

घर के आगे/NN_NST एक पेड़ था।

तुम मेरे बाद/NN_NST आए थे।

Here is the explanation (with some modifications to suit here) that appears in the standards document for the conclusion drawn above for the category of NST:
आगे Age (front), *पीछे pIche*(behind), *ऊपर Upara*(above/upstairs), *नीचे nIce*(below/down), *बाद bAda*(after),*पहले pahale*(before), *अंदर andara*(inside), *बाहर bAhara (outside)* etc.

The tag N_NST has been included to cover an important phenomenon of Indian languages.

Certain expressions such as ऊपर, नीचे, पहले, आगे etc. are content words denoting time and space. These expressions, however, are used in various ways.

These words often occur as temporal or spatial arguments of a verb in a given sentence taking the appropriate *vibhakti* (case marker):

Ex2. वह ऊपर सो रहा था
vaha ***Upara*** *so rahA thA* .
'he' 'upstairs' 'sleep' 'PROG' 'was'
"He was sleeping upstairs".

Ex3. वह पहले से कमरे में बैठा था
vaha ***pahale se*** *kamare meM bEThA thA.*
'he' 'beforehand' 'from' ' room' 'in' 'sitting' 'was'
"He was sitting in the room from beforehand"

Ex4. *तुम बाहर बैठो*
tuma ***bAhara*** *bETho*
'you' 'outside' 'sit'
"You sit outside".

Apart from functioning like an argument of a verb, these elements also modify another noun taking postposition 'kA'.

Ex5. उसका बड़ा भाई ऊपर के हिस्से में रहता है
usakA baDZA bhAI ***Upara ke hisse meM*** *rahatA hE*
'his' 'elder' 'brother' 'upstairs' 'of' 'portion' 'in' 'live' 'PRES'
"His elder brother lives in the upper portion of the house".

Apart from occurring as a nominal expression, they also occur as a part of a postposition along with 'ke'. For example,

Ex6. घड़े के ऊपर थाली रखी है
ghaDZe ***ke Upara*** *thAlI rakhI hE.*
'pot' 'of' 'above' 'plate' 'kept' 'is'
The plate is kept on the pot".

Ex7. तुम घर के बाहर बैठो
tuma ghara ke ***bAhara*** *bETho*
'you' 'home' 'of' 'outside' 'sit'
"You sit outside the house".

'Upara' and '*bAhara*' are parts of complex postpositions *'ke Upara'* and '*ke bAhara*' in (ex5) and (ex6) respectively which can be translated into English prepositions 'on' and 'outside'.

For tagging such words, one possible option is to tag them according to their syntactic

function in the given context. For example in (ex6) above, the word '*Upara*' is occurring as part of a postposition or a relation marker. It can, therefore, be marked as a postposition. Similarly, in (ex2) and (ex5) above, it is a noun, therefore, mark it as a noun and so on. Alternatively, since these words are more like nouns, as is evident from above they can be tagged as nouns in all their occurrences. The same would apply to '*bAhAra'* (outside) in examples (ex3), (ex4) and (ex7).
However, if we follow any of the above approaches we miss out on the fact that this class of words is slightly different from other nouns. These are nouns which indicate 'location' or 'time'. At the same time, they also function as postpositions in certain contexts. Moreover, such words, if tagged according to their syntactic function, will hamper machine learning. Considering their special status, it was considered whether to introduce a new tag, NST, for such expressions. The following five possibilities were discussed:

a) Tag both (ex4) & (ex7) as NN
b) Tag both (ex4) & (ex7) as NST
c) Tag (ex4) as NN & (ex7) as NST
d) Tag (ex4) as NST & (ex7) as PSP
e) Tag (ex4) as NN & (ex7) as PSP

After considering all the above, the decision was taken in favour of (b). The decision was primarily based on the following observations:
(i) *'bAhara'* in both (ex4) and (ex8) denotes the same expression (place expression 'outside')
(ii) In both (ex4) and (ex7), *'bAhara'* can take a vibhakti like a noun (**bAhara ko** bETho, ghara **ke bAhara ko** bETho)
(iii) If a single tag is kept for both the usages, the decision making for annotators would also be easier.

Therefore, a new tag **NST** is introduced for such expressions. The tag **NST** will be used for a finite set of such words in any language. For example, Hindi has
आगे Age (front), *पीछे pIche*(behind), *ऊपर Upara*(above/upstairs), *नीचे nIce*(below/down), *बाद bAda*(after),*पहले pahale*(before), *अंदर andara*(inside), *बाहर bAhara (outside)* etc.

2. Pronouns (PR)

The category of pronoun has been divided into six sub-categories. These include personal, reflexive, relative, reciprocal, wh-word and indefinite. These categories should be self-explanatory and follows the same definitions as posited in common linguistic literature.

2.1 Personal Pronouns (PR_PRP)

Personal pronouns cover all the pronouns that denote to person, place or thing. This includes the all their cases as well for example: मैं, हम, मेरा, हमारा, मुझे, हमें, मुझी, हमीं, तुम, तुम्हारा, तुम्हें, तुझी etc.

2.2 Reflexive Pronoun (PR_PRF)

Reflexive pronouns are the ones that denote to ownership to its antecedent which can be either a noun or a pronoun. The only examples of reflexive pronouns in Hindi are the अपना/अपने/अपनी, स्वयं and खुद. (Please report to me if you find any other word that can go into this category) E.g.:

वह अपना/PR_PRF खाना खुद/PR_PRF पकाता है।

राम स्वयं/PR_PRF सामर्थ्यवान है।

वह अपने/PR_PRF गांव गया है।

2.3 Relative Pronouns (PR_PRL)

The relative pronouns are those pronouns whose antecedent can be either a noun or a pronoun. However, these pronouns do not make any difference in number or gender as in the case of personal pronouns. The relative pronoun in Hindi is represented by जो and its inflected forms e.g.

जो/PR_PRL करता है वो भरता है।

वह लड़का जो/PR_PRL आया था चला गया।

जिसे/PR_PRL आपने बुलाया था वह नहीं आया।

The list of relative pronouns may be exhaustive. One has to make sure that one makes the appropriate difference between the relative pronoun and the relative demonstrative pronoun. Note that in the following sentences the occurrences of जो and जिस are demonstrative relatives (DM_DMR) and not relative pronouns (PR_PRL)

जो/DM_DMR लड़का यहां आया था वह चला गया।

जिस/DM_DMR लड़के को आपने बुलाया था वह नहीं आया।

2.4 Reciprocal Pronoun (PR_PRC)

Reciprocal pronouns denote some reciprocity. This is commonly denoted by आपस में, परस्पर

वे लोग आपस/PR_PRC में झगड़ा कर रहे थे।

परस्पर/PR_PRC बात करके सुलझा लीजिए।

2.5 Wh-word Pronouns (PR_PRQ)

The wh-word pronouns are typically the pronouns that are used to ask questions. These words are कब, क्या, कौन, कहां, कैसे etc.

क्या/PR_PRQ लाए हो?

क्या/PR_PRQ तुमने खा लिया?

फिर कब/PR_PRQ आओगे?

2.6 Indefinite Pronouns (PR_PRI)

The indefinite pronouns refer to unspecified objects, places or things. These words are किसी, कोई, कहीं, कभी etc.

शरीर का कोई/ PR_PRI हिस्सा खराब हो गया हो

The only difference between PR_PRI (pronoun indefinite) and DM_DMI (demonstrative indefinite) is pronoun indefinite has some previous reference, शरीर का कोई/ PR_PRI हिस्सा, referring to one of the parts of body.

3. Demonstratives (DM)

The category of demonstrative has been separated from the category of pronouns as the demonstratives mainly indicate about a noun and does not act as anaphora. The demonstratives have been sub-categorized into four divisions- deictic, relative, wh-words and indefinite.

3.1.Deictic (DM_DMD)

The deictic demonstratives are default demonstratives that demonstrate the noun it modifies. The deictic demonstratives in Hindi are typically यह, वह, ये, and वे. These always occur before the noun they modify.

यह/DM_DMD शहर बहुत प्राचीन है।

उस/DM_DMD घर की छत पक्की है।

3.2.Relative Demonstrative (DM_DMR)

The relative demonstrative occur in the same form as the relative pronoun. The difference is only that these relatives are always followed by a noun that it modifies.

जिस/DM_DMR गांव में मैं गया था वह बहुत सुंदर था।

जो//DM_DMR नहर टूट गई थी उसकी मरम्मत की जा रही है |

3.3.Wh-Word Demonstrative (DM_DMQ)

The wh-demonstratives are the same wh-words (or question words) which act as wh-pronouns. The difference is that in their demonstrative function they do not ask question, rather only demonstrates. The wh-word demonstratives are कोई, किसी, कौन etc.

कोई/DM_DMQ लड़का आया था।

यह बात किसी/DM_DMQ से छुपी नहीं है।

जाने कौन/DM_DMQ यहां रहता है।

जाने कहां/DM_DMQ से आई है।

3.4 Indefinite Demonstratives (DM_DMI)

Like for indefinite pronouns, the indefinite demonstratives refer to unspecified objects, places or things. These words are किसी, कोई, कहीं, कभी etc.

किसी/DM_DMI दिन वह आएगा।

कोइ/DM_DMI लडका आया था।

4.Verb (V)

The verb in the BIS has been divided into two categories- main and auxiliary. The dual distinction is to decipher main verb and the auxiliary verb(s) in any sentence. While the auxiliary verb is a closed set of verb, the main verb can be anything from a root verb to any of its inflected forms. Each sentence or clause must have a main verb. A sentence can have one more auxiliary verbs. As in the following sentences:

Simple Verbal occurrences

मैं/PR_PRP मोहन/N_NNP हूं/V_VM ।/RD_PUNC

वह/PR_PRP पढ़ता/V_VM है/V_VAUX ।/RD_PUNC

वह/PR_PRP पढ़/V_VM रहा/V_VAUX है/V_VAUX ।/RD_PUNC

वह/PR_PRP पढ़/V_VM रहा/V_VAUX होगा/V_VAUX ।/RD_PUNC

Compound Verbs

मैं/PR_PRP अब/RB स्वयं/PR_PRF समझ/V_VM सकता/V_VAUX हूं/V_VAUX ।/RD_PUNC

उसने/PR_PRP किताब/N_NNC फाड़/V_VM दी/V_VAUX है/V_VAUX ।/RD_PUNC

उसने/PR_PRP मुझे/PR_PRP नहीं/RP_NEG जाने/V_VM दिया/V_VAUX ।/RD_PUNC

Conjunct Verbs

The conjunct verbs are preceded by a noun or an adjective. This preceding noun or adjective are marked with their own category while the verb (which is usually a set of verbs called sometimes light verb or vector verbs) coming after, it will be marked as the main verb.

वह/PR_PRP गुस्से/N_NNC से/PSP लाल/JJ हो/V_VM रहा/V_VAUX था/V_VAUX ।/RD_PUNC

उसने/PR_PRP तुम्हें/PR_PRP मूर्ख/JJ बना/V_VM दिया/V_VAUX ।/RD_PUNC

Verbal Nouns/ Gerunds

In the BIS standards document, the category of verbal nouns or gerund is not considered for Hindi or other Indo-Aryan languages. Such verbs function as the noun in the sentences but are capable of taking their own arguments. For this very reason, they are to be marked as the verbs. Even though they do not function as the main verb in the sentence, they are to be marked with the label of main verb.

खेलना/V_VM स्वास्थ्य/N_NNC के/PSP लिए/PSP अच्छा/JJ होता/V_VM है/V_VAUX ।/RD_PUNC

हंसने/N_NNC से/PSP अवसाद/N_NNC कम/JJ होता/V_VM है/V_VAUX ।/RD_PUNC

मैं/PR_PRP आपसे/PR_PRP मिलने/V_VM को/PSP उत्सुक/JJ हूं/V_VM ।/RD_PUNC

Nouns Derived from Verbs

Nouns derived from verbs (such as पढ़ाई, खुदाई, पुताई, etc.) are to be marked as nouns. These are marked as nouns because they cannot take an argument of their own in the sentence.

सुबह/N_NN की/PSP पढ़ाई/N_NN याद/N_NN रहती/V_VM है/V_VAUX ।/RD_PUNCT

*सुबह की किताब पढ़ाई याद रहती है।

खुदाई/N_NN का/PSP काम/N_NN चल/V_VM रहा/V_VAUX है/V_VAUX ।/RD_PUNCT

*गढ्ढा खुदाई का काम चल रहा है।

Participial constructions of verbs acting as modifiers

The standards document does not specify anything about the participial and the –कर constructions which also abound in Hindi. I presume that these would also be taken as main verbs.

दौड़ता/V_VM हुआ/V_VAUX लड़का/N_NN गिर/V_VM पड़ा/V-VAUX ।/RD_PUNCT

सैनिक/N_NN दौड़कर/V_VM आया/V_VM ।/RD_PUNCT

दौड़ने/V_VM वाला/V_VAUX लड़का/N_NN प्रथम/QT_QTO आया/V_VM ।/RD_PUNC

4.1 Finite Verbs (V_VM_VF)

There is a separate tag for finite verb, limited to specific languages. Example from Punjabi:

hhd4 ਇਹ\DM_DMD ਦੰਦਾਂ\N_NN ਨੂੰ\PSP ਗੰਦਾ\JJ ਅਤੇ\CC_CCD ਸਾਹਾਂ\N_NN ਨੂੰ\PSP ਬਦਬੂਦਾਰ\N_NN ਬਣਾ\V_VM_VNF ਦਿੰਦੇ\V_VM_VF ਹਨ\V_VAUX |\RD_PUNC

4.2 Non-finite verb (V_VM_VNF)

This tag is also for limited languages to tag the non-finite feature of the verb. Consider the following Punjabi example:

hhd5: ਇੱਥੇ\N_NST ਦਿੱਤੇ\V_VM_VNF ਕੁੱਝ\QT_QTF ਅਸਾਨ\JJ ਨੁਸਖਿਆਂ\N_NN ਦੀ\PSP ਮਦਦ\N_NN ਨਾਲ\PSP ਤੁਸੀਂ\PR_PRP ਆਪਣੇ\PR_PRF ਦੰਦਾਂ\N_NN ਨੂੰ\PSP ਸਾਫ਼\JJ ਅਤੇ\CC_CCD ਸਾਹਾਂ\N_NN ਨੂੰ\PSP ਤਾਜ਼ਾ\JJ ਰੱਖ\V_VM_VNF ਸਕਦੇ\V_VAUX ਹੋ\V_VAUX |\RD_PUNC

4.3 Infinitive (V_VM_VINF)

Infinitives are often preceded by 'to'; but not necessarily. There are Indian languages in which the demarcation between infinitival verb and gerundial verb is blurred. Therefore, this tag is also limited for some specific languages in which this demarcation could be easily observed. Example from Punjabi:

hhd21:ਦੰਦਚਿਕਿਤਸਕ\N_NN ਤੋਂ\PSP ਦੰਦਾਂ\N_NN ਦੀ\PSP ਜਾਂਚ\N_NN ਨਿਯਮਿਤ\JJ ਰੂਪ\N_NN ਨਾਲ\PSP ਕਰਾਓ\V_VM_VINF |\RD_PUNC

4.4 Gerunds (V_VM_VNG)

In the BIS standards document, the category of verbal nouns or gerund is not considered for Hindi or other Indo-Aryan languages. Such verbs function as the noun in the sentences but are capable of taking their own arguments. For this very reason, they are to be marked as the verbs. Example from Telugu:

htd11019:చోఖీడాణీ\N_NNP చుట్టుపక్కల\RB ప్రతిచోటా\N_NN కట్టబడిన\V_VM_VNF వేదికల\N_NN మీద\PSP మరొక్క\JJ చోట\N_NN కాల్‌బెలియా\N_NNP మరొక్క\JJ చోట\N_NN గుండ్రంగా\RB తిరగడం**V_VM_VNG** అయితే\RP_RPD మరొక్క\JJ చోట\N_NN అల్గోజా\N_NNP నృత్యం\N_NN చేస్తూ\V_VM_VNF జానపద\N_NN కళాకారులుగా\RB కనిపిస్తారు\V_VM_VF .\RD_PUNC

5. Adjective (JJ)

Adjective has not been sub-divided into any categories. There is one category for an adjective which is self-explanatory. These are mostly attributive adjectives. For quantifiers, there is a separate category defined. Examples: बड़ा, छोटा, लाल, सुंदर, चर, अचर, अग्रिम etc.

However, they pose a problem in cases like बहता झरना (flowing waterfall), *घिरी दीवारें* (surrounded walls) etc. The word बहता (flowing) shows some activity in itself; e.g. in the sentence

(v) पानी बहता रहता है
"Water keeps flowing."

this word would be tagged as verb. But it becomes a challenge to tag it in cases like *बहता झरना,* as it would qualify for the adjective tag (JJ). In cases like these, the word *बहता* is tagged as Verb.

6. Adverb (RB)

Adverb also is mono-category part-of-speech. The standards document says that the category of adverb (RB) is only for manner adverbs. For example, words like धीरे, जल्दी, तेज, etc. are adverbs.

However, there are many other words which qualify for adverbial category, but are not given this category. So, some confusion remains. For the example the words like बिल्कुल, एकदम, हमेशा, इसीलिए, and अक्सर are such words which mostly modify the verb or the whole of a clause and are not manner adverbs.

Due to the absence of any category within (manner) Adverb, we include these words like बिल्कुल, एकदम, हमेशा under this category.

7. Postposition (PSP)

Postpositions are all the parts-of-speech that work as case marker. Words like में, का, के, की, ने, पर, etc. are examples of postposition.

राम का/PSP भाई सुंदर है।

गांव में/PSP लोग कुएं का/PSP पानी पीते हैं।

There are postpositions in Hindi and other Indian languages that are formed with more than words contributing to it. Such postpositions are called complex postpositions. Examples of such postpositions are के लिए, के ऊपर, के बाद, के पीछे etc. Such complex postpositions are to be marked as PSP for all the words that constitute it. Example:

मैंने राम के/PSP लिए/PSP यह पुस्तक लाई है।

मेज के/PSP ऊपर/PSP पुस्तक रखी है।

मेरे बाद/PSP राम आएगा।

8. Conjunction (CC)

Conjunctions are non-content words that act as joiners of phrases or clauses within a sentence. The category of conjunction has been divided into two sub-categories of coordinator and subordinator.

8.1. Coordinators (CC_CCD)

Coordinators are typically the words that join two phrases, of the same category. It can join one or more of a noun phrase, a verb phrase or a clause. The common examples are और, तथा, व, बल्कि, किन्तु, परन्तु etc.

राम और/CC_CCD श्याम घर जा रहे हैं।

उसने ऐसा केवल कहा ही नहीं बल्कि/CC_CCD किया भी।

8.2.Subordinators (CC_CCS)

Subordinator conjunctions typically conjoin two clauses and the second clause is subordinated. That is the clause conjoined by the subordinator word is the subordinate clause against the main clause. Typical subordinator words in Hindi are अगर, क्योंकि, तो, कि, चूंकि etc.

अगर/CC_CCS आज वह आया तो/CC_CCS मैं उसे बताऊंगा।

उसने कहा कि/CC_CCS वह आज नहीं आएगा।

We include words like इसीलिए under this category as it joins the subordinating clause with the main one.

8.3 Quotative (CC_CCS_UT)

A **quotative** is grammatical device to mark reported speech in some languages. It can be equated with "spoken quotation marks". Tamil Example:

hhd8007 ஒரு\QT_QTC வேளை\N_NN குழந்தை\N_NN அதிகமாக\RB அழுதால்\N_NN உடனே\PSP குழந்தைக்கு\N_NN மூக்கு\N_NN அடைப்பாக\RB இருக்கிறதா\V_VM_VNF அல்லது\CC_CCD அதன்\PR_PRP காதில்\N_NN ஏதேனும்\PR_PRQ இருக்கிறதா\V_VM_VNF என்று\CC_CCS_UT பார்க்க\V_VM_VINF வேண்டும்\V_VM_VF .\RD_PUNC

9. Particles (RP)

Particles are words that do not decline and also do not fall into any other categories described above and elsewhere. Four sub-categories - default, interjection, intensifier and negation - have been created to cover the particles in Hindi.

9.1. Particle Default (RP_RPD)

The default particles are ही, तो and भी.

वह आने ही/RP_RPD वाला है।

वह पढ़ ही तो/RP_RPD रहा था।

जाने भी/RP_RPD दो अब।

9.2 Classifier (RP_CL)

A **classifier** sometimes called a **measure word** is a word or morpheme used in some languages to classify the referent of a countable noun according to its meaning. In languages that have classifiers, they are often used when the noun is being counted or specified (i.e., when it appears with a numeral or a demonstrative). It is very common phenomenon in Bangla:

hhd359 ভারতে\N_NNP আট\QT_QTC জনের\RP_CL মধ্যে\PSP একটি\QT_QTC ব্যাক্তি\N_NN নিজের\PR_PRF বয়সের\N_NN যে\PR_PRL কোনও\DM_DMQ সময়ে\N_NN ক্যানসার\N_NN দ্বারা\PSP আক্রান্ত\N_NN হতে\V_VM_VNF পারে\V_VM_VF ।\RD_PUNC

9.3 Interjection (RP_INJ)

Interjections are particles which denote exclamation utterances. The common exclamatory marks in Hindi are ओह, आह, हाय, उफ, अरे, हे, ओ, अबे, etc.

ओह/RP_INJ ये क्या हो गया!

उफ//RP_INJ हद हो गई यार!

9.4 Intensifier (RP_INTF)

Intensifiers are words that intensify the adjectives or adverbs. The common intensifiers in Hindi are बेहद, बहुत, अत्यंत, etc.

वह बेहद/RP_INT कमजोर हो चला है।

वह बहुत/RP_INT ही धीरे चलता है।

यह अत्यंत/RP_INT मार्मिक कहानी है।

9.5 Negation (RP_NEG)

The negation particles are the words that indicate negation. These include नहीं, न, ना, मत, बिना etc.

आज वह नहीं/RP_NEG आया।

शोर ना/RP_NEG करना।

10 Quantifiers (QT)

Quantifiers are the words that modify nouns or adjectives and indicate quantity. These have been sub-categorized into three parts- general, cardinals and ordinals.

10.1 General (QT_QTF)

The general quantifiers do not indicate any precise quantity, e.g. थोड़ा, बहुत, ज्यादा, कम, कुछ, etc.

10.2 Cardinals (QT_QTC)

The cardinal quantifiers are absolute numbers, either in digits or in words such as 1, 2, 3, एक, दो, तीन etc.

10.3 Ordinals (QT_QTO)

The ordinals denote the order part of the digits such as पहला, दूसरा, तीसरा etc. The ordinals in Hindi also inflect for gender and number and take the oblique case marker, thus पाँचवीं, बीसवां, तीसवें etc.

11 Residuals (RD)

The category of residuals has been demarcated for the words that are usually not intrinsic part of the language/speech. Divided into five parts, these include foreign words, symbols, punctuations, unknown words and echo-words.

11.1 Foreign Words (RD_RDF)

The foreign words are all the words that are not written in the Devanagari script.

11.2 Symbols (RD_SYM)

The symbols are the characters that are not part of the regular Devanagari script such as *, @, #, $, % etc.

11.3 Punctuations (RD_PUNC)

Punctuations include the characters that are considered as the regular punctuation marks in Hindi, e.g. (,),,,!,?,- etc.

11.4 Unknown (RD_UNK)

Unknown words would the words for which a category cannot be decided by the annotator. These may include words from phrases or sentences from a foreign language written in Devanagari.

11.5 Echo-Words(RD_ECH)

The echo-words are the words that are formed by the morphological process known as echo-formation e.g. पानी-वानी, खाना-वाना etc

---

Report on discussions on the Guidelines for Annotation – ILCI (Phase 1)

ILCI-ILMT POS Workshop

IIIT Hyderabad 21-23 July, 2012

The following edits were made to the Annotation guidelines after discussions between the language groups PIs and Prof Dipti Misra at IIIT, Hyderabad. The language group members are required to go through this document carefully and revert back in case of any problem. They should take note of the changes/discussions made for their language specifically and incorporate the changes within the corpus already tagged by them in Phase 1.

1. Noun:

1.1 Common

-In Kashmiri and Urdu, some compound nouns are separated by a white space and the first part of the word does not exist independently in the language. In such cases, it is difficult to mark the POS category of the first word. As of now, the problem is being dealt with by putting an underscore (zero space marker) between the two words and marking it as a whole word.

-*Wala* in Hindi is treated as a noun when it occurs with words like *sabziwala*, *doodhwala* etc. However, when it occurs with a space, e.g. *sabzi wala*, it is marked as a PSP (postposition).

-In Indian languages, some general confusion may persist on whether a word should be tagged as Noun or Adjective, as most adjectives also occur as nouns. In such cases, the behavior of the word must be taken into account.

Example: *goro ne bharat par 200 saal raj kiya.*

The (white skinned) foreigners ruled over India for 200 years.

In this case, the word *goro* should be tagged as a noun, although it is an adjective, *gora* (white). This is so because here, the word inflects for case, which is a property of nouns.

-In languages like Malayalam, nouns may agglutinate with morphemes of other categories, e.g. Noun + Postposition + Aux. In such cases, the whole word is to be marked as a Noun, without segmentation. In Bangla, case may sometimes be affixed to the noun without any space marker. In those cases also, the word has to be marked as Noun.

1.2 Proper

This tag should be applied specifically to entities which have a unique referent. Names of cities, countries, people, institutions, organizations, specified month (in a date) etc are proper nouns. Examples: Paris, India, Abraham Lincoln, Aligarh Muslim University, Rashtriya Sanskrit Sansthan, 26th January 1987, Bay of Bengal. Nouns specifying class, objects etc must be marked common noun (N_NN). Names of medicines, diseases, seasons, flowers, animals, months (unless specified) are all N_NN.

Example: lions, dog, rose, crocin, cancer, February etc.

When there is confusion, we have to decide whether the name is that of a class or instance.

Example:

Dog → NN

↓

Greyhound, Bulldog, Alsatian →NN

↓

Timmy, Jim, Max → NNP

Properties of NNP:

1. In most cases, it will have to be transliterated. Example: Jawaharlal Nehru University (exceptional cases: BangAl kI khADI for Bay of Bengal)
2. It will refer to a unique individual entity. Example: Gandhi.
3. In a lot of cases, it is context-dependent. Example: He is a *Shakespeare*\N_NN; Roshni chali gayi. (Roshni may refer to a girl's name or to electricity)

1.3 Verbal

This category has been deleted as it is not required for any language, as of now.

1.4 Spatio-temporal

Nouns which denote time and space are tagged as spatio-temporal nouns. This category was introduced because words like *upar, neeche, bahar* etc can occur as nouns as well as part of complex postpositions and create ambiguities during tagging. These nouns are to be tagged as NST in all cases. Example: *ke upar, ke andar, neeche se* etc.

However, not all words denoting time or space are to be marked as NST. Only nouns that denote time and space which participate in both complex PSPs and NNs. Example: *ke upar, se neeche* etc.
In cases like *mera ateet* (my past), *mera kal* (my future) the words *ateet* and *kal* will be marked as common nouns.

*All groups must create a list of NST words in their language and include it within the guidelines.

2. Pronouns

They differ from the Demonstrative category in the respect that Demonstratives are used as pointers towards following noun, whereas Pronouns act in place of nouns.

Example: *Vah gaya* (He went) → *vah* here refers to someone and is used in the place of a noun or to address a previously mentioned noun.

Vah ladka mera bhai hai. (That boy is my brother) →In this case, the *vah* occurs before a noun and is used to indicate that specific noun, therefore will be marked as Demonstrative.

3. Demonstratives

See above category.

4. Verb

4.1 Main

4.1.1 Finite

The decision to disambiguate the finiteness is language dependent. In some languages finite verbs inflect for tense. In languages like Gujarati, tense is not marked. In Manipuri, mood defines finiteness.

In agglutinative languages, the word may have morphemes of different categories. For example: in Malayalam, a word may be of the form

Noun + Postposition + Verb Finite + Verb Non Finite + Adverb

In such cases, the word needs to be segmented first, if possible and then tagged.

For Punjabi, at POS level, verbs will not be marked for Finite\ Non- Finite category. They should be marked with the Verb_Main tag.

4.1.2 Non-finite

See above category

4.2 Auxiliary

Auxiliary should be marked only when it is separately lexicalized. If it is part of an inflectional suffix, then it does not need to be separately marked. A word must not be segmented if the purpose is only to mark the auxiliary. Refer section 1.1 (Common Noun)

5. Adjective

Most adjectives also occur as nouns, so some confusion may persist.

Example: गोरों ने राज किया।

It inflects so it is a noun here. We have to see its behavior and then tag. Example: मेरा कल|

Adjective vs. Verb:

In cases like *rotA bachchA*, *rotA* will be tagged as Adjective in all cases.

In cases like *rotA huA bachchA*, *rotA* is tagged as Verb_Main and *huA* as Verb_Aux.

6. Adverb

The decision to make it only manner adverbs will be edited in the next version.

In Assamese and Kashmiri, in certain cases, reduplication of adjectives makes the phrase an adverb, as the adverbial marker is absent in this language. In those cases, the words will be tagged depending on the context. If it functions as an Adverb, it should be marked accordingly.

7. Postposition

In Hindi, *5 va* will be marked as PSP. In *doodh vala*, if *vala* occurs separately, it will be marked PSP.

8. Conjunction

No issues.

9. Particle

9.1 Default

Mark something as a default particle when it is a particle and it does not come under any of the other given sub-categories.

9.2 Interjection

In a phrase like *hei ram*! although the whole phrase is an interjection, we mark it as *hei*\RP_INJ *ram*\N_NNP

9.3 Intensifier

In cases like *bohot sundar ladkiyan*, although *bohot* is a quantifier at the lexical level, it acts as an intensifier in this context, intensifying the quality of the adjective.

9.4 Negation

This should be marked when the particle has the semantics of complete negation. Example: *nahi, na* in Hindi. However, the Hindi *bina* does not imply negation, but exclusion. It should therefore be marked as a default particle.

10. Quantifiers

Quantifiers are generally not considered as a grammatical category and may function as nouns or adjectives. However, lexical items which denote quantity should be marked as quantifiers. Example: *kuch, thoda, ek, teesra* etc in Hindi.

11. Residual

11.1 Foreign word

Only those words which are written in a foreign script should be marked as foreign word, even if the annotator understands the foreign script. Example: If a word written in Devanagari script appears in Konkani text (written in Romi script), then the Devanagari word is a Residual Foreign Word.

11.3 Punctuation

This category needs a separate tag as it may sometimes denote a semantic meaning. If the meaning is dependent on the placement of a hyphen, e.g. classifier in Bangla, then the word should be segmented and tagged accordingly.

Example: *chata* (to lick) and *cha-ta* (the tea)

The first word should be tagged as V_VM and the second as *cha*\N_NN -\RD_PUNC *ta*\RP_CL

11.4 Unknown

A word will be marked for Unknown category when the annotator cannot decide the POS category that the word belongs to.

11.5 Echo Word

Reduplicated words, that is, words which occur in repetition, should be marked according to their respective lexical category.

Example: *dheere-dheere* will be marked as *dheere*\RB -\RD_PUNC *dheere*\RB.

Reduplication can occur before or after the noun, depending upon the language.

In some languages, like Telugu, partial reduplication occurs. In some other cases, like in Kahmiri, lexical items which do not occur separately when reduplicated, have meaning. Example: *vilvil*

However, phrases like *pani-vani, chai-vai* etc contain echo words which do not belong to any POS category. In such cases, the word which belongs to a POS category should be marked with that tag and the echo word should be marked as Residual Echo Word. Example: *pani*\N_NN -\PUNC *vani*\RD_ECH

In some languages like Malayalam, the echo words may occur before the noun. Example: *kovitta kovna, patta pagalu* etc. For those cases, a new tag ECH_B (Echo_Before) will be introduced in the next version.

Glossing abbreviations

The system of interlinear glossing as practiced in this volume is based on the *Leipzig Glossing Rules* developed by members of the Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig.

Punctuation

Parallel in the object-language/transliteration and the meta-language/gloss:

- Connects **segmentable** morphemes.

= Marks **clitic** boundaries.

~ Indicates **reduplication** morphemes.

Isolated in the object-language/transliteration or the meta-language/gloss:

. Separates two meta-language elements that correspond to a single object-language element (that **cannot be separated** into corresponding different morphemes).

: Separates two meta-language elements that correspond to a single object-language element that **could be separated** into corresponding different morphemes, **but the author chose not to separate it**; *or* it separates two object-language elements that correspond to a single meta-language element that could be separated into corresponding different morphemes, but the author chose not to separate it.

_ **Combines** two meta-language elements that correspond to a single object-language element; *or* it combines two unconnected object-language elements that correspond to a single meta-language element.

\ Marks a grammatical property in the object-language signaled by a morpho-phonological change (***ablaut***, **mutation**, **tone alternation**, etc.)

Different meanings in the meta-language/gloss and the object-language transliteration:

… in the meta-language:

( ) Marks **inherent**, non-overt categories.

[ ] Marks a property that **does not correspond to an overt element** in the object-language.

… in the object-language/philological transliteration:

( ) Marks scholarly **reconstruction of non-overt phonemes** in writing systems with ‘defective’ spellings, e.g. unwritten glides in Egyptian.

⸢ ⸣ Marks a passage in the written object-language that is **partially destroyed**.

[ ] Marks a passage in the written object-language that is **completely destroyed**; eventually it contains scholarly reconstructions of the destroyed passage.

{ } Marks **extra elements** in the object-language that are to be classified either as **scribal errors** (philological emendation) or as **redundant information** as part of certain orthographical conventions .

‹ › Marks additions to the object-language data, whose **missing** is to be classified as **scribal errors** (philological emendation).

* Marks **reconstructed** forms.

Abbreviations

1	first person
2	second person
3	third person
ABL	ablative
ABSTR	abstractum
ACC	accusative
ACT	active
ADJ	adjective
ADJZ	adjectivizer/adjectivization
ADV	adverb(ial)
ADVZ	adverbializer/adverbialization
AGR	agreement
AGT	agent marker
ALL	allative
ANT	anterior
AOR	aorist
APPL	applicative
ART	article
ATTD	attached
ATTN	attention catching particle
BASE	particle as base for enclitic pronoun
BEN	benefactive
C	communis
CAUS	causative
CIRC	circumferential
CJVB	conjunctional verb
CL	(written) classifier (trad. ‘determinative’)
CNSV	consecutive particle or suffix
COLL	collective
COND	conditional
CONN	connective particle
COP	copula
CORD	coordinating particle
DAT	dative
DATLOC	dative-locative
DEF	definite
DEM	demonstrative
DIST	distal
DISTR	distributive
DU	dual
ELAT	elative
EXCL	exclusive
EXCLM	exclamative
EXTR	exterior
F	feminine
FUT	future
GEN	genitive
GN	god’s name
GRND	ground
ILL	illative
IMP	imperative
IMPRS	impersonal
INDF	indefinite
INESS	inessive
INF	infinitive
INFR	inferior
INS	instrumental
IPFV	imperfective
IPRF	imperfect
LOC	locative
LOCADV	locative-adverbialis
M	masculine
MED	medium
MOD	modal
MODP	modal particle
MP	medio-passive
N	neuter
N…	non- (e.g.npst non-past)
NEG	negation, negative
NINFL	not inflected (here: for gender and number)
NMLZ	nominalizer/nominalization
NOM	nominative
OBJ	object
OBLV	obligative
OBP	*Ortsbezugspartikel*
OPT	optative
ORD	ordinal number
PASS	passive
PERS	personal
PFV	perfective
PL	plural
PLUPRF	plu-perfect
PN	personal name
POSS	possessive
PP	adpositional phrase

PREP	preposition	SBRD	subordinating particle
PRF	perfect	SG	singular
PROH	prohibitive	SP	sentence particle
PROX	proximal/proximate	STAT	stative
PRS	present	STC	*status constructus*
PRT	preterite (= PST past)	STPR	*status pronominalis*
PST	past	SUPR	superior
PTCL	particle	TA	tense/aspect gram
PTCP	participle	TAM	tense/aspect/mode gram
QUOT	quotative	TERM	terminative-adverbialis
RECP	reciprocal	THMZ	thematizer/thematization
REFL	reflexive	TN	toponym
REL	relative	TOP	topic
RES	resultative	TR	transitive
SBJN	subjunction	VENT	ventive
SBJV	subjunctive	VOC	vocative

Bibliography

Di Biase-Dyson, Kammerzell & Werning 2009
Camilla Di Biase-Dyson, Frank Kammerzell & Daniel A. Werning, Glossing Ancient Egyptian: Suggestions for Adapting the Leipzig Glossing Rules, in: *Lingua Aegyptia. Journal of Egyptian Language Studies* 17 (2009), 243–266; <http://www.gwdg.de/~dwernin/published/DiBiase_Kammerzell_Werning-2009-Glossing_Ancient_Egyptian.pdf>.

Glossing Ancient Languages
Glossing Ancient Languages, Open access Wiki, http://wikis.hu-berlin.de/interlinear_glossing/, edited by Daniel A. Werning, Berlin: Humboldt University Berlin.

Leipzig Glossing Rules
The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-Morpheme Glosses, ed. by the Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology (Bernard Comrie, Martin Haspelmath) and by the Department of Linguistics of the University of Leipzig (Balthasar Bickel); <http://www.eva.mpg.de/lingua/resources/glossing-rules.php>, <http://www.eva.mpg.de/lingua/pdf/LGR08.02.05.pdf>, Leipzig, Sept., 12th 2008 (download 2009).

April 3, 2007

Unified style sheet for linguistics

These guidelines grew out of discussions among a group of editors of linguistics journals during 2005-2006 and were approved on January 7, 2007. They are intended as a "default, but with discretion to use common sense", to quote David Denison on the matter. Our principles, as elaborated primarily by Stan Dubinsky, are:

1. **Superfluous font-styles should be omitted.** Do not use small caps for author/editor names, since they do not help to distinguish these from any other bits of information in the citation. In contrast, italics are worthwhile for distinguishing volume (book, journal, dissertation) titles [+ital] from article and chapter titles [-ital].

2. **Superfluous punctuation should be left out.** Once italic is adopted to distinguish volumes from articles/chapters (as above), then single or double quotations around article titles are superfluous and only add visual clutter.

3. **Differing capitalization styles should be used to make category distinctions.** Use capitalization of all lexical words for journal titles and capitalize only the first word (plus proper names and the first word after a colon) for book/dissertation titles and article/chapter titles. This is a useful diagnostic for discriminating between titles that are recurring and those that are not. The journal style for capitalization should also be applied to the title of book series. Thus, the citation of a *SNLLT* volume would be punctuated: *Objects and other subjects: Grammatical functions, functional categories, and configurationality* (Studies in Natural Language and Linguistic Theory 52).

4. **All author/editor first names should be spelled out.** Not doing so only serves to make the citation less informative. Without full first names, the 20th century index for *Language* alone would conflate five different people as ‘J. Smith’, four as ‘J. Harris’, three each under ‘A. Cohen’ and ‘P. Lee’, two each under ‘R. Kent’, ‘J. Anderson’, ‘H. Klein’ and ‘J. Klein’.

5. **The ampersand is useful.** Use ampersand to distinguish higher and lower order conjuncts, i.e. [W & X] and [Y & Z], as in Culicover & Wilkins and Koster & May. It is relatively easy to see that reference is made here to two pairs of authors here (cf. Culicover and Wilkins and Koster and May).

6. **Name repetitions are good.** While using a line ____ may save a little space, or a few characters, it also makes each such citation referentially dependent on an antecedent, and the effort of calculating such antecedents is more than what it saved typographically. Each citation should be internally complete.

7. **Four digit year plus period only.** Extra parentheses are visual clutter and superfluous.

8. **Commas and periods and other punctuation.** Separate citation components with periods (e.g., Author. Year. Title.) and subcomponents with commas (e.g., Author1,

Author2 & Author3). Please note the ampersand (&), rather than the word “and” before the name of the last author, and no comma before the “&”. The use of the colon between title and subtitle and between place and publisher is traditional, but we do not use it between journal volumenumber and pagenumbers.

9. Parentheses around ed. makes sense. Commas and periods should be used exclusively to separate citation components (e.g., "Author. Year."), or subcomponents (e.g. "author1, author2 & author3). Since "ed." is neither a component nor a subcomponent, but a modifier of a component, it should not be separated from the name by a comma:

surname, firstname = author
surname, firstname (ed.). = editor (NOT surname, firstname, ed.)
surname, firstname & firstname surname (eds.) = editors

10. For conference proceedings, working papers, etc. For conference proceedings published with an ISSN, treat the proceedings as a journal: Include both the full conference name and any commonly used acronym for the conference (*BLS, WCCFL*, etc.) in the journal title position. For proceedings not published with an ISSN, treat the proceedings as any other book, using the full title as listed on the front cover or title page. If the title (and subtitle if there is one) only includes an acronym for the conference name, expand the acronym in square brackets or parentheses following the acronym. If the title does not include an acronym which is commonly used for the conference name, include the acronym in square brackets or parentheses following the conference name. The advantage of including the acronym after the society title is that it makes the entry much more identifiable in a list of references.

11. Use “edn.” as an abbreviation for “edition”, thus “2nd edn.”. This avoids ambiguity and confusion with “ed.” (editor).

12. Names with “von”, “van”, “de”, etc. If the "van" (or the "de" or other patronymic) is lower case and separated from the rest by a space (e.g. Elly van Gelderen), then alphabetize by the first upper-case element:

Gelderen, Elly van

The addition of "see ..." in comprehensive indices and lists might be helpful for clarification:

van Gelderen, Elly (see Gelderen)

13. Names with “Jr.”, “IV.”, etc. Following library practice, list elements such as “Jr.” as a subelement after names, separated by a comma.

Smith, Sean, Jr.

14. Use "In" to designate chapters in collections. This makes the book's format maximally similar to the standard citation format. This, in turn, would be time-saving when the author or the editor notice that more than one article is cited from a given collection and hence that that book's details should be set out as a separate entry in the references (and the full details deleted from the articles' entries).

author. year. chaptertitle. In editorname (ed.), collectiontitle, pagenumbers. publisher.

15. Journal volume numbers. We favor: volumenumber(volumeissue). startingpage-endingpage. Thus: 22(1). 135-169. Note the space between volume number/issue and page numbers. Special formatting (e.g., bold for volume number) is superfluous. Issue numbers are a parenthetical modifier (cf. "ed." above) of the volume number. While it is not NECESSARY information for identifying the article, it is extremely USEFUL information.

16. Dissertations/theses. These conform to the already-widespread Place: Publisher format and fit readily into the rest of the standard: Cambridge, MA: MIT dissertation. Instead of archaic state abbreviations, use the official two-letter postal abbreviations. Note that national and other traditions vary in exactly what is labeled 'thesis' versus 'dissertation' and in distinguishing 'PhD' from 'doctoral' dissertations.

> Cambridge, MA: MIT dissertation.
> Chapel Hill: UNC MA thesis.

17. On-line materials. The basic information here — author, date, title — remains the same, and the URL where the resource was found takes the place of publisher or journal. We urge authors to include the date the material was accessed, in parentheses after the URL, since new versions often replace old ones. For a .pdf file, this would be the date of downloading, but for a resource like an on-line dictionary consulted repeatedly, a range of dates may be needed. For additional discussion of handling on-line citations, authors may want to consult this guide:

> Walker, Janice R. & Todd Taylor. 1998. *The Columbia Guide to Online Style*. New York: Columbia University Press.

Example references

Blevins, Juliette. 2004. *Evolutionary phonology.* Cambridge: Cambridge University Press.

Casali, Roderic F. 1998. Predicting ATR activity. *Chicago Linguistic Society* (*CLS*) 34(1). 55-68.

Chomsky, Noam. 1986. *Knowledge of language*. New York: Praeger.

Coetsem, Frans van. 2000. *A general and unified theory of the transmission process in language contact.* Heidelberg: Winter.

Franks, Steven. 2005. Bulgarian clitics are positioned in the syntax. http://www.cogs.indiana.edu/people/homepages/franks/Bg_clitics_remark_dense.pdf (17 May, 2006.)

Iverson, Gregory K. 1983. Korean /s/. *Journal of Phonetics* 11. 191-200.

Joseph Salmons 5/12/05 12:14 PM
Comment: No use of caps/small caps for author/editor names.

Joseph Salmons 5/12/05 12:34 PM
Comment: Capitalize only the first word of a book/dissertation title (plus proper names and the first word after a colon).

Joseph Salmons 6/1/05 2:26 PM
Comment: Call conference proceedings titles according to the name of the society, including the meeting's acronym in parentheses. Do not include "proceedings of the" or "papers from the".

Joseph Salmons 6/6/05 2:13 PM
Comment: Author/editor first names should be spelled out (recommended, but not obligatory).

Joseph Salmons 5/12/05 12:21 PM
Comment: Four digit year plus period only; no parentheses.

Joseph Salmons 6/6/05 2:13 PM
Comment: alphabetize names with "von", "van", "de", etc. according to first upper-case element.

Joseph Salmons 4/1/07 10:28 AM
Comment: For on-line materials, give the date the resource was accessed.

Joseph Salmons 5/15/05 11:51 AM
Comment: No italics with article titles.

Joseph Salmons 5/15/05 11:56 AM
Comment: Capitalize all lexical words in journal or series titles.

Iverson, Gregory K. 1989. On the category supralaryngeal. *Phonology* 6. 285-303.

Johnson, Kyle, Mark Baker & Ian Roberts. 1989. Passive arguments raised. *Linguistic Inquiry* 20. 219-251.

Lahiri, Aditi (ed.). 2000. *Analogy, leveling, markedness: Principles of change in phonology and morphology* (Trends in Linguistics 127). Berlin: Mouton de Gruyter.

McCarthy, John J. & Alan S. Prince. 1999. Prosodic morphology. In John A. Goldsmith (ed.), *Phonological theory: The essential readings*, 238-288. Malden, MA & Oxford: Blackwell.

Murray, Robert W. & Theo Vennemann. 1983. Sound change and syllable structure in Germanic phonology. *Language* 59(3). 514-528.

Oxford English Dictionary, 2nd edn. 1989. Oxford: Oxford University Press.

Pedersen, Johan. 2005. The Spanish impersonal *se*-construction: Constructional variation and change. *Constructions* 1, http://www.constructions-online.de. (3 April, 2007.)

Rissanen, Matti. 1999. Syntax. In Roger Lass (ed.), *Cambridge History of the English Language*, vol. 3, 187-331. Cambridge & New York: Cambridge University Press.

Stewart, Thomas W., Jr. 2000. *Mutation as morphology: Bases, stems, and shapes in Scottish Gaelic.* Columbus, OH: The Ohio State University dissertation.

Webelhuth, Gert (ed.). 1995. *Government and binding theory and the minimalist program: Principles and parameters in syntactic theory.* Oxford: Blackwell.

Yu, Alan C. L. 2003. *The morphology and phonology of infixation.* Berkeley, CA: University of California dissertation.

Joseph Salmons 5/15/05 11:50 AM
Comment: Repeat names for each entry.

Joseph Salmons 5/12/05 12:23 PM
Comment: Comma used to separate subcomponents (e.g. author1, author2, author3.)

Joseph Salmons 5/12/05 12:24 PM
Comment: Period used to separate citation components (e.g. author. year. title.)

Joseph Salmons 5/16/05 7:04 AM
Comment: Give series information in parentheses, capitalizing all lexical words.

Joseph Salmons 5/12/05 12:35 PM
Comment: Use ampersand (also for in-text reference).

Joseph Salmons 5/16/05 5:18 PM
Comment: Capitalize only first word of article and chapter titles, plus first word after colon .

Joseph Salmons 6/1/05 1:50 PM
Comment: Use "In" for chapters in collections.

Joseph Salmons 5/12/05 12:26 PM
Comment: Volume (book, journal, dissertation) title italicized.

Joseph Salmons 5/12/05 12:29 PM
Comment: Lexical words in journal titles capitalized.

Joseph Salmons 5/12/05 12:29 PM
Comment: Format for journal information: volumenumber(volumeissue).startingpage-endingpage, e.g. 22(1).135-169.

Joseph Salmons 6/6/05 1:37 PM
Comment: Use "edn." as an abbreviation for "edition"

Joseph Salmons 4/3/07 8:19 AM
Comment: For on-line journal, give journal URL after title (and volume, if applicable), followed by date consulted.

Joseph Salmons 6/6/05 2:40 PM
Comment: Place "Jr.", "IV", etc. after names.

Joseph Salmons 5/12/05 12:30 PM
Comment: Use parentheses around "ed."; do not separate from last name by comma.

Joseph Salmons 5/12/05 12:30 PM
Comment: Use two-letter postal abbreviations.

Joseph Salmons 5/16/05 6:52 AM
Comment: Format for dissertations/theses: City, State: Institution dissertation/MA thesis (e.g. Cambridge, MA: MIT dissertation; Chapel Hill, NC: UNC MA thesis).

LANGUAGE: JOURNAL OF THE LINGUISTIC SOCIETY OF AMERICA

Language **Style Sheet**

This style sheet results from the accumulated wisdom of those people who have participated in the editing of *Language* over the years, and have worked to establish and maintain consistency in formatting in the journal's publications. Please note that this style sheet does **not** need to be followed in the preparation of manuscripts that are being submitted to the journal for review. Its purpose is to guide authors whose papers have been accepted for publication in the final preparation of their manuscripts for typesetting. Manuscripts that depart from the style sheet will be returned to the author for corrections in egregious cases.

Important note about file formats: If at all possible, please prepare **the text** of your manuscript in a basic word-processing program like Word and submit it as a .doc/.docx/.rtf file. Trees, AMVs, or anything else that requires formatting that is difficult in such programs can be submitted in other formats as special matter (see below for details). Please note that our typesetting process does **not** use camera-ready text.

If you work in LaTeX and submit your manuscript as a .tex file, our typesetters will charge us to convert it to .doc, which is required for both us and them to work with the file and typeset your article. If you must submit it in .tex, please send us **all style files** that were used in conjunction with the .tex file (.sty, .bib, etc.).

In all cases, please send a **.pdf file** of your manuscript along with the other files, for our reference.

1. BASIC FORMATTING

a. Set paper size to Letter, 8½ x 11.

b. Set line spacing to 1.5 throughout the document.

c. Use extra space between sections.

d. Use 12 point font throughout the document (**including** title, headings, and notes), in a simple roman face except where indicated below (§3).

e. Set margins of 1 inch (2.54 cm.) on all four sides of the paper.

f. Left-align throughout the document (do not justify).

g. Do not use line-end hyphens.

h. Use a single space after all punctuation, not two spaces.

i. Number all pages of the entire manuscript serially in the upper right corner.

j. Do not use any other headers or footers.

k. Special matter (tables, tableaux, figures, maps) should be given on separate pages at the end of the document, or in a separate file or files (see §2 below for details about the preparation of special matter).

l. Use endnotes rather than footnotes, numbered with arabic numerals.

m. The LSA urges contributors to *Language* to be sensitive to the social implications of language choice and to seek wording free of discriminatory overtones. In particular, contributors are asked to follow the LSA Guidelines for Nonsexist Usage, originally published in the December 1996 *LSA Bulletin*, and available online at: http://lsadc.org/info/lsa-res-usage.cfm.

n. Use the following order and numbering of pages.

i. page 0: title and subtitle; authors' names and affiliations as they will appear at the beginning of the article; email addresses for all authors (and mailing addresses, as desired, for first or all authors), to appear at the end of the article.
ii. page 1: title and subtitle only
iii. page 2: abstract of about 100 words (for articles and short reports) with asterisked acknowledgment footnote (footnote placeholder should come at end of abstract) and **a list of 5–7 keywords** (place after the abstract: *Keywords*: X, Y …).
iv. body of the work
v. (appendix, if applicable)
vi. references, beginning on a new page
vii. notes, beginning on a new page
viii. all special matter (or in separate file or files; see below)

2. SPECIAL MATTER

Special matter includes all tables, tableaux, figures, trees and other diagrams, and art work (not example sentences, rules, or formulas).

a. Numbering of special matter

i. Tables should be numbered separately from other examples: Table 1, Table 2, etc.

ii. Figures (including charts, graphs, pictures, trees) should be numbered separately from other examples and tables: Figure 1, Figure 2, etc.

iii. OT tableaux and some syntactic trees can be numbered as regular examples within the text, but should still follow the conventions outlined below.

b. Key each piece of special matter to its proper place in the body of the manuscript with a notation of the following sort on a separate line in the manuscript.

<INSERT FIGURE 1 ABOUT HERE>

<INSERT TABLE 5 ABOUT HERE>

For a tableau or other special matter that is numbered as a regular example, include the example number, followed by the notation:

(15) <INSERT Tableau 15 HERE>

or with a legend:

(15) Tableau illustrating ranked bigram constraints

<INSERT Tableau 15 HERE>

c. File formats: Tables, OT tableaux, and other text-based special matter (including some figures) should be set in a word-processing program, and submitted in a .doc file or the equivalent.

Each table, tableau, or other text-based special matter should appear on a separate page at the end of the main text file, or on a separate page in a separate file of special matter. Centered below each table or figure, put its number, followed by a brief legend.

TABLE 1. Basic ordering typology for adjacent affixes.

A tableau or other special matter that is numbered as a regular example does not need a legend, but should be keyed to its place in the text.

<Tableau for example 15>

d. Figures that are not text-based should be sent as individual files (containing just the figure itself, not including the figure number and legend); these files can be sent in various formats, such .pdf, .eps, .jpg, .bmp, .xls, .doc, depending on how the figure was originally created and what would give the best product. Figures should be as high resolution as possible, and should be in black and white. **Name** figure files according to their number (Figure1, Figure2b, Figure 2b, etc.). The figures in these files should be camera-ready.

In addition to the separate figure files, figure numbers and legends should appear on a separate page at the end of the main text file, or in a separate file of special matter; images of the figures can be included in that file as well, for reference.

FIGURE 2. Average pre-*ka* placement for 172 roots.

The accompanying .pdf file of the whole document that is sent should also include all of the figures and tables with their legends. Please note, however, that the figures cannot be set from this file or from an image inserted into a .doc file, and thus it is important to send a separate file for each individual figure, as indicated above.

3. TYPEFACES AND SPECIAL FONTS

a. Use *italics* for all cited linguistic forms and examples in the text. Do **not** use italics for emphasis, or to mark common loanwords or technical terms: ad hoc, façon de parler, ursprachlich, binyan, etc.

b. Use SMALL CAPITALS to mark a technical term at its first use or definition, or to give emphasis to a word or phrase in the text.

c. Please do not capitalize names of laws, theories, or hypotheses; the first appearance may be given in small capitals to indicate the use as a technical term.

d. Use **boldface** for certain forms in Oscan and Umbrian, and to distinguish Gaulish and other forms originally written in the Greek alphabet.

e. **Boldface** can also be used to draw the reader's attention to particular aspects of a linguistic example, whether given within the text or as a numbered example.

f. If special fonts are required, as much as possible use unicode-based fonts. For phonetic fonts, we prefer Doulos SIL or Doulos SIL Compact (available from sil.org). If you use any other fonts, please send the fonts along with your other files.

g. If you have **any** problems getting a particular symbol to show up correctly in the manuscript you send us, please include a note with the file clearly explaining what the symbol should look like, with a picture for reference. Please do not just try to approximate the correct symbol by adjusting spacing, font size, etc., since those formatting details will be lost in the regular typesetting process.

4. PUNCTUATION

a. Use single quotation marks, except for quotes within quotes. The second member of a pair of quotation marks should precede any other adjacent mark of punctuation, unless the other mark is a necessary part of the quoted matter: The word means 'cart', not 'horse'. He asked, 'What can we hypothesize about this example?'.

b. Do not enclose any cited linguistic examples in quotation marks. See §6.

c. Indent long quotations (more than about forty words) without quotation marks.

d. Do not hyphenate words containing prefixes unless a misreading will result (e.g. *nonlinguistic*, *postvocalic*, etc.); hyphenate if the stem begins with a capital letter: non-Dravidian, Proto-Athabaskan.

e. Indicate ellipsis by three periods, close set, with a blank space before and after, like … this.

f. Use a comma before the last member of a series of three or more coordinate elements: A, B, and C; X, Y, or Z (the 'Oxford comma'). Do not use a comma after the expressions e.g. and i.e.

g. Use a period (full stop) before numbered examples, tables, or figures, not a colon.

5. NOTES

a. Number all notes to the body of the text serially throughout the document.

b. The note reference number in the body of the text is a raised arabic numeral, not enclosed in parentheses. Place note numbers at the ends of sentences wherever possible, or after a comma, semicolon, or other punctuation mark that indicates a pause or natural break; the note reference number should be placed **after** the punctuation mark. Do not link more than one note to a single place in the text.

c. All notes should be placed at the end of the text (following the references) as endnotes, 1.5 spaced, 12 point font, like the rest of the text (see §1).

d. Each note should be a separate paragraph beginning with its reference number, raised above the line and not followed by any punctuation mark.

e. Place the acknowledgment footnote at the end of the abstract, keyed with an asterisk.

f. Number footnotes to special matter (numbered as a, b, c) separately for each piece of special matter and place them as footnotes on the same page as the special matter.

6. CITED FORMS

a. Do not italicize numbered examples. Italicize words or other linguistic forms only when cited within the text.

b. Enclose transcriptions either within (phonetic) square brackets or within (phonemic) slashes: the suffix [q], the word /rek/. Do not italicize bracketed transcriptions.

c. Use angle brackets for specific reference to graphemes: the letter <q>.

d. Transliterate or transcribe all forms in any language not normally written with the Latin alphabet, including Greek, unless there is a compelling reason for using the original orthography. Use IPA symbols unless there is another standard system for the language.

e. After the first occurrence of non-English forms, provide a gloss in single quotation marks: Latin *ovis* ‘sheep’ is a noun. No comma precedes the gloss and no comma follows, unless necessary for other reasons: Latin *ovis* ‘sheep’, *canis* ‘dog’, and *equus* ‘horse’ are nouns. See §8 for other instructions on glosses.

7. NUMBERED EXAMPLES, RULES, AND FORMULAS

a. Place each numbered item on a separate line with the number in parentheses; indent after the number; use lowercase letters to group sets of related items.

(2) a. Down the hill rolled the baby carriage.
    b. Out of the house strolled my mother’s best friend.

b. In the text, refer to numbered items as 2, 2a, 2a,b, 2a-c (with no parentheses).

c. Examples in notes should be numbered as (i), (ii), (iii), etc., and should be referred to as such in the text.

8. GLOSSES AND TRANSLATIONS OF EXAMPLES

Examples not in English must be translated or glossed as appropriate. Sometimes, both a translation and a word-for-word or morpheme-by-morpheme gloss are appropriate.

a. Place the translation or gloss of an example sentence or phrase on a new line below the example, indented.

(26) La nouvelle constitution approuvéé, le président renforça ses pouvoirs.
'The new constitution approved, the president consolidated his power.'

b. Align word-for-word or morpheme-by-morpheme glosses of example phrases or sentences with the beginning of each original word; use tabs to make alignments rather than multiple spaces.

(17) Omdat duidelijk is dat hie ziek is.
because clear is that he ill is

c. Observe the following conventions in morpheme-by-morpheme glosses:

i. Place a hyphen between morphs within words in the original, where relevant, and a corresponding hyphen in the gloss; do **not** use any hyphens in the gloss that do not have corresponding hyphens in the original.

ii. If one morph in the original corresponds to two or more elements in the gloss (cumulative exponence), separate the latter by a period, except for persons; there is no period at the end of a word.

(4) siastr-yn-y malunk-i
sister-POSS-M.PL.NOM picture-M.PL.NOM
'the sister's pictures'

iii. Gloss lexical roots in lowercase roman type.
Gloss persons as 1, 2, 3, and 4.
Gloss all other grammatical categories in small capitals.

iv. Abbreviate glosses for grammatical categories. List the abbreviations in a note.

9. ABBREVIATIONS

a. Abbreviations ending in a small letter have a following period; abbreviations ending in a capital do not.

b. Abbreviations such as e.g., i.e., etc., cf., and others should only be used within parentheses; elsewhere, spell out ‘for example, … ’, ‘that is, … ’, and so forth.

c. Names of languages used as adjectives are often abbreviated prenominally; the editors follow the practice of Merriam-Webster dictionaries for these abbreviations.

d. Use prime notation (e.g. S', V'') rather than bar notation.

10. SECTION HEADINGS

a. Use the same roman typesize as the body of the text for all headings.

b. The number and the following period should be in **boldface**; the heading text should be in SMALL CAPITALS.

c. Capitalize only the first word of the heading.

d. Do not use more than two levels of headings: for example, **1** or **2.3** are fine, but not **3.2.4**. If a further division of the section is necessary, simply use SMALL CAPS for the subsection heading, with no number.

> METHODS. Experiment 1 took place in a sound-attentuated lab …

e. Place section headings on a line with the section number and the first line of the section.

> **1.** INTRODUCTION. The recent renaissance of …

11. CITATIONS IN THE TEXT

Within the text, give only a brief citation in parentheses consisting of the author's surname, the year of publication, and page number(s) where relevant: (Rice 1989, Yip 1991:75–76).

a. If the citation is of the **work**, place either everything within parentheses: (e.g. Joseph & Janda 2004:121), or nothing in parentheses: More discussion of issues related to historical reconstruction can be found in Joseph & Janda 2004:121. In this case, use an ampersand between two authors’ names, and if there are more than two authors, use the surname of the first author, followed by et al.: (see Yip et al. 1995).

b. If, by contrast, the citation is of the **author**, and the author's name is part of the text, then use this form: Rice (1989:167) comments that ..., Joseph and Janda (2004:121) note that …, Yip and colleagues (1995:34) illustrate this … . Please note the following specifications: only the date (and page numbers) are in parentheses; use ‘and’ rather than ampersand between two author names; use ‘and colleagues’ or the like rather than ‘et al.’ for more than two authors.

c. Do not use notes for citations only, other than for website URLs when necessary.

12. REFERENCES

At the end of the manuscript, provide a full bibliography, 1.5 spaced, beginning on a separate page with the heading REFERENCES.

a. Arrange the entries alphabetically by surnames of authors, with each entry as a separate hanging indented paragraph. Surnames with a separately written prefix (e.g. von, de, van der, etc.) should be alphabetized by the prefix.

van der Sandt, Rob A. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics* 9.333–77.

Wilson, Deirdre. 1975. *Presuppositions and non-truth-conditional semantics*. London: Academic Press.

b. List multiple works by the same author in ascending chronological order. No distinction should be made between works for which the author was the editor vs. the author.

Hymes, Dell H. 1974a. *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.

Hymes, Dell H. (ed.) 1974b. *Studies in the history of linguistics: Traditions and paradigms*. Bloomington: Indiana University Press.

Hymes, Dell H. 1980. *Language in education: Ethnolinguistic essays*. Washington, DC: Center for Applied Linguistics.

c. Use suffixed letters a, b, c, etc. to distinguish more than one item published by a single author in the same year.

d. Do not replace given names with initials unless the person always uses initials: Dixon, R. M. W., but Lehiste, Ilse.

e. Use a middle name or initial only if the author normally does so: Heath, Shirley Brice; Oehrle, Richard T.

f. Author names should be given in small capitals (if you cannot easily set small capitals, please leave them in regular font—do **not** set them as all capitals and/or in a smaller font size).

g. Please give each reference a date; we do not list works with ‘to appear’, ‘in progress’, ‘in press’, etc. in lieu of a date. If the reference has been accepted for publication, list it with the estimated date of publication, and include ‘to appear’ at the end of the entry. If the reference has not yet been accepted for publication, please give it the date corresponding to the version you referenced, and list it as a manuscript, with the author’s place and affiliation (see the Miner 1990 entry below).

Sprouse, Jon; Matt Wagers; and Colin Phillips. 2011. A test of the relation between working memory capacity and syntactic island effects. *Language*, to appear.

h. If more than two articles are cited from the same book, list the book as a separate entry under the editor's name, with cross-references to the book in the entries for each article;

similarly, if a book is cited independently within the text and references, individual articles from that book should cross-reference the book.

BUTT, MIRIAM, and WILHELM GEUDER (eds.) 1998. *The projection of arguments: Lexical and compositional factors*. Stanford, CA: CSLI Publications.

CROFT, WILLIAM. 1998. Event structure in argument linking. In Butt & Geuder, 21–63.

i. Book and journal names should be given in italics. Capitalize only the first word of the title and subtitle of an article or book, as well as any other words required to be capitalized in the language's orthography.

j. Each entry should contain the following elements in the order and punctuation given: (first) author's surname, given name(s) or initial(s); given name and surname of other authors. Year of publication. Full title and subtitle of the work. For a journal article: Full name of the journal and volume number (roman type).inclusive page numbers for the entire article. For an article in a book: title of the book, ed. by [full name(s) of editor(s)], inclusive page numbers. For books and monographs, the edition, volume or part number (if applicable); series title (if any) in parentheses. Place of publication: Publisher.

k. Use en-dashes between page numbers; include appropriate page numbers as follows: 12–17, 143–46, 198–205, 1147–55, 1195–203, etc.

l. If a reference is published online—for example, an unpublished manuscript hosted on the author's website, or an open-access online publication, such as a journal or conference proceedings—please include a link to the article, as in the examples below. Do not include links for articles published in hard-copy books or journals, unless the electronic version is open-access and hosted by the owner of the copyright.

DONOHUE, MARK. 2009. Geography is more robust than linguistics. *Science* e-letter, 13 August 2009. Online: http://www.sciencemag.org/cgi/eletters/324/5926/464-c.

SALTZMAN, ELLIOT; HOSUNG NAM; JELENA KRIVOKAPIC; and LOUIS GOLDSTEIN. 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. *Proceedings of the 4th International Conference on Speech Prosody (Speech Prosody 2008)*, Campinas, 175–84. Online: http://aune.lpl.univ-aix.fr/~sprosig/sp2008/papers/3inv.pdf.

SUNDELL, TIMOTHY R. 2009. Metalinguistic disagreement. Ann Arbor: University of Michigan, MS. Online: http://faculty.wcas.northwestern.edu/~trs341/papers.html.

m. Additional examples are given below.

DORIAN, NANCY C. (ed.) 1989. *Investigating obsolescence*. Cambridge: Cambridge University Press.

GROPEN, JESS; STEVEN PINKER; MICHELLE HOLLANDER; RICHARD GOLDBERG; and RONALD WILSON. 1989. The learnability and acquisition of the dative alternation in English. *Language* 65.203–57.

HALE, KENNETH, and JOSIE WHITE EAGLE. 1980. A preliminary metrical account of Winnebago accent. *International Journal of American Linguistics* 46.117–32.

MINER, KENNETH. 1990. Winnebago accent: The rest of the data. Lawrence: University of Kansas, MS.

PERLMUTTER, DAVID M. 1978. Impersonal passives and the unaccusative hypothesis. *Berkeley Linguistics Society* 4.157–89.

POSER, WILLIAM. 1984. *The phonetics and phonology of tone and intonation in Japanese*. Cambridge, MA: MIT dissertation.

PRINCE, ELLEN. 1991. Relative clauses, resumptive pronouns, and kind-sentences. Paper presented at the annual meeting of the Linguistic Society of America, Chicago.

RICE, KEREN. 1989. *A grammar of Slave*. Berlin: Mouton de Gruyter.

SINGLER, JOHN VICTOR. 1992. Review of *Melanesian English and the Oceanic substrate*, by Roger M. Keesing. *Language* 68.176–82.

STOCKWELL, ROBERT P. 1993. Obituary of Dwight L. Bolinger. *Language* 69.99–112.

TIERSMA, PETER M. 1993. Linguistic issues in the law. *Language* 69.113–37.

YIP, MOIRA. 1991. Coronals, consonant clusters, and the coda condition. *The special status of coronals: Internal and external evidence*, ed. by Carole Paradis and Jean-Francois Prunet, 61–78. San Diego, CA: Academic Press.

Data and language documentation

JEFF GOOD
University at Buffalo

1. INTRODUCTION

The topic of this chapter is the relationship between data and language documentation. Unlike many fields of study, concerns regarding data collection and manipulation play a central role in our understanding of, and theorizing about, language documentation. The field to a large extent, in fact, owes its existence to a shift in focus in the goals of linguistic field work from concerns regarding outputs derived from primary data, like grammars and dictionaries, to the collection of the primary data itself.

When trying to understand the role of data in language documentation, the first question we must consider is what precisely do we mean by *data*? Beginning with the work of Himmelmann (1998), it has become customary in language documentation to distinguish between *primary data*—constituting recordings, notes on recordings, and transcriptions—and *analytical resources*—like descriptive grammars and dictionaries—constructed on the basis of, and via generalization over, primary data. While making this conceptual distinction is essential to the practice and theorizing of language documentation, most individuals or teams working on language documentation projects are ultimately interested in both collecting primary data and producing the kinds of analytical resources associated with traditional language description, most prominently grammars, dictionaries, and texts (whether oriented for community or academic use). Therefore, each will be considered here. That is, the discussion will cover topics both regarding the collection, storage, and manipulation of primary data as well as the mobilization (see Holton (this volume)) of that data to create analytical resources. While it is also important to keep in mind that *data* is not synonymous with *digital data*, for the most part, in this chapter, only digital data will be discussed. Generally, digital, rather than analog, data has been the focus of work in language documentation both because new data is typically captured solely in digital form at present and because analog data is increasingly being digitized so that it can be manipulated and disseminated with digital tools. Discussion of important aspects of digitization—i.e., the process through which a digital representation of a non-digital object is created—can be found in the E-MELD School of Best Practices in Digital Language Documentation[2] (Boynton et al. (2006)), and an exemplary case study of the digitization process can be found in Simons et al. (2007).

This chapter will focus on conceptual issues rather than specific technical recommendations, though such recommendations may be discussed to provide illustrative examples. This is because our understanding of the conceptual issues evolves at a much slower rate than the technical recommendations, which change as the technologies we use for capturing and analyzing data themselves change and, therefore, largely outpace the speed through which works like this one make their way into publication. At least for the time being, the best way to find answers to questions like *What audio recording device should I use?* or *What*

[2] http://e-meld.org/school/

software should I use for text annotation? will be to use online resources like the E-MELD School just mentioned above, electronic publications like *Language Documentation and Conservation*[4] or the *Transient Languages and Cultures* blog[5], and email lists like the one run by the Resource Network for Linguistic Diversity[6]. The role of a chapter like this one is, therefore, not so much to tell language documenters what to do but, rather, to put issues surrounding data in a broader context, to allow them to understand why recommendations take a particular shape, and to better equip them to evaluate new technologies as they become available. Readers looking to augment the discussion here with more specific recommendations will find Austin (2006) helpful, as it covers similar subject matter to this chapter but on a more concrete level. More advanced conceptual discussion can be found in Bird and Simons (2003) which overlaps partially with the discussion here but also goes beyond it in many respects.

This chapter divides the discussion into the following topics: Data types in section 2, data structures in section 3, data formats in section 4, metadata in section 5, a brief discussion of needs assessment in section 6, and a concluding section on the linguist's responsibilities for navigating the relationship between their data and new technologies in section 7.

2. DATA TYPES

The discussion in this section is subdivided here into the topics of recordings, transcriptions, and traditional descriptive resources, each of which is treated in turn, followed by discussion of community-oriented versus academic-oriented data. I do not treat written language, as opposed to transcription, specifically here both because of the general emphasis in language documentation on collecting instances of spoken language (though, see Woodbury (this volume)) and because, from a data management perspective, written representations do not generally differ significantly from transcription. I also do not discuss scanned images, though these can play a role in language documentation, as well, particularly for projects making use of paper-based materials. (See Simons et al. (2007) for a relevant case study using scanned images to create high-quality documentary resources.)

2.1 Recordings

In the present context, following Himmelmann (1998:162), primary data will be used to refer two very distinct classes of resources. Direct recordings of events on the one hand, and written representations of those events on the other. Direct recordings include, most prominently, audio recordings and, increasingly, video recordings as well as photographs, though they can also include more "exotic" resources like laryngographs or palatograms. These kinds of resources are sometimes referred to as *raw data* (see, for example, Schultze-Berndt 2006:215), to highlight the fact that they can be created without extensive linguistic analysis, unlike transcriptions.

However, one should not be complacent and assume that the "rawness" of this data implies that it represents a purely objective rendering of a given

[4] http://nflrc.hawaii.edu/ldc/

[5] http://blogs.usyd.edu.au/elac/

[6] http://rnld.org/

communicative event. All recording involves selection: what to record, when to record, how to record, etc. And these selections, made by a person, not a machine, can shape the record tremendously, not only influencing the perceived quality of the recording but also emphasizing and deemphasizing features of the recorded event and the language in possibly significant ways. For example, use of a unidirectional microphone in making an audio recording will result in a resource where one speaker is framed as more central to a speech event than any others, while use of an omnidirectional microphone will produce a resource where different participants' voices are recorded more equally. Analytical linguistic factors may influence which kind of microphone is chosen for a given recording. In a grammatical elicitation session with a single speaker, for example, a unidirectional microphone is more likely be chosen, while for a recording made of a story an omnidirectional microphone may be used even though only one participant has the special role of storyteller if the story is being told in a society where audience participation is the norm. Similar issues arise in making the choice to make video recordings in addition to audio ones. For certain kinds of events—or even languages, in the case of sign languages—use of video may be essential, but the question of what visual aspects of a scene to capture is a particularly clear kind of selection.

Therefore, while the production of raw recordings involves less intensive linguistic analysis than creating, say, a transcription, it should not be forgotten that it involves a series of choices, some of which may be mostly pragmatic in nature (e.g., not to use a video recorder for a given session to conserve scarce battery power) while others (e.g., not to use a video recorder because a session is deemed to be visually "uninteresting") may actually be informed by an underlying—if only implicit—theory of recording. This point bears special importance for researchers choosing to adopt collaborative modes of fieldwork with their communities (see, e.g., Mithun 2001, Grinevald 2003, Dwyer 2006 for relevant discussion) or who intend their work to assist in community language maintenance and revitalization projects (see, e.g., Mosel 2006, Nathan 2006 and Hinton, Penfield, McCarty and Coronel-Molina (this volume)) since community input may be required to ensure that the form of the recordings is not unduly skewed towards research needs.

2.2 Transcriptions

Transcriptions (often annotated—see Schultze-Berndt (2006) for detailed discussion of annotation) have generally been treated under the heading of primary data due to the fact that they are intended to be a representation of a particular speech event rather than serving as generalizations over distinct speech events. Unlike recordings, however, the creation of transcriptions implies extensive linguistic analysis (see, e.g., Himmelmann 2006), and they, therefore, occupy a territory between documentation and description. (The same could be said for written representations of language in general, though in some cases written examples of language serve as primary data not merely by convention but because they constitute the only available representation of a given use of language.)

A crucial difference between transcriptions and recordings, however, is that recording techniques and technologies tend to be general in nature while transcription is a specifically linguistic task. The devices used by linguists to make audio recordings are more or less the same as those used by musicians, oral historians, journalists, etc. However, many of the transcription conventions used by linguists, e.g., the International Phonetic Alphabet or aligned glossing, are domain-specific and largely under the control of the linguistic community.

An important consequence of this is that while language documenters will generally be reactive in the domain of recording techniques, they will often need to be proactive in the domain of transcription techniques. Thus, language documentation work is at the forefront of the next generation of transcription and annotation tools, as evidenced, for example, by the ELAN annotation tool[7] (see Berez (2007) for a review) produced specifically in the context of the Dokumentation Bedrohter Sprachen (DoBeS) Programme.

2.3 Descriptive resources

Three kinds of resources have long been given a special place in descriptive linguistics: texts, dictionaries, and grammars. If the most important feature distinguishing descriptive resources from documentary resources is the fact that they attempt to arrive at generalizations about a language based on raw data, it is clear that texts are less prototypically descriptive than dictionaries and grammars. However, to the extent that they are normalized and edited for internal consistency, they shift from being records of a specific speech event, as with a transcription, to being representations of an idealized speech event and, therefore, begin to cross the boundary into description.

By contrast, dictionaries and grammars are unambiguously instances of description. A dictionary is an attempt to generalize over the known lexical items of a language to create a concise summary of their uses and meanings, while a grammar generalizes over textual and elicited data to create a summary of the phonological, morphological, and syntactic constructions of a language. Formal work making use of extensive language data is not generally construed as an essential part of the creation of an adequate description of a language. However, in the present context, it could, in principle, also be included under the broad heading of language description as well. In practice, however, the field of linguistics tends to reserve the term for informal description rather than formal description. (See Dryer (2006) for discussion of relevant issues.)

2.4 Community data versus academic data

It has become standard practice for linguists documenting under-resourced languages to consider ways in which their work can result in outputs not only for use in academic spheres, but also community ones. Accordingly, brief discussion of this issue is in order here.

It is important to be clear that trying to serve multiple communities will always require more work than serving only one community. At the same time, modern technology can significantly reduce the extra burden placed on language documenters who opt to do this. This is because, digital data, unlike data on paper, can be copied and transformed relatively easily. To take an example outside of the domain of language documentation, it has now become commonplace for individuals to transform text documents from whatever format they were originally composed in (e.g., in the native format of their word processing program) to Portable Document Format (PDF), a format specifically designed to create documents which are readable across a wide range of computer platforms.

[7] http://www.lat-mpi.eu/tools/elan/

This transformation process has been largely automated requiring only a trivial investment of time on the part of the user.

The kinds of data transformations required to allow a single language resource to serve speaker and research communities, of course, will never be as straightforwardly automated as conversion to PDF if for no other reason than the fact that groups interested in such functionality do not have the economic power to attract the interest of large software companies. However, as will be discussed in the following sections, if the data collected by a project is encoded in certain ways, allowing it to serve multiple audiences becomes more manageable. Furthermore, if non-proprietary, open formats are used and the way the data is encoded is well-documented (see section 4), anyone with sufficient technical expertise will be able transform the original data into new formats, substantially increasing the potential impact of a project and perhaps also decreasing the workload of the language documenter who would not, then, be required to perform such data transformations themselves.[9]

3. DATA STRUCTURE VERSUS IMPLEMENTATION

Often, when people talk about their data, they conflate the abstract structure of the various datatypes they collect with the ways those datatypes happen to be encoded in a particular *view*—that is, a way of representing the data in a human-readable form. Thus, for example linguists often speak of *interlinear glossed text* as a basic data type when, in reality, it is probably better understood as a specific way of expressing a data type we might refer to *morphologically-analyzed text*—that is, a text on which an exhaustive morphological analysis has been performed. Interlinear glossing has become widely adopted as an effective way of presenting such a morphological analysis, in particular on the printed page, but it is just one of many imaginable ways of doing this. For example, in early twentieth century texts one sometimes finds a convention where individual words are associated with endnotes giving analytical details well beyond what is possible with a short gloss (see, for example, the texts in Boas (1911)). And, of course, using modern hypertext methods, interactive forms of glossing have become possible as well.

Linguists tend to think of interlinear glossed text as a basic data type in and of itself because it represents a primary way they interact with texts, and, this is, of course, a perfectly natural conflation. However, when it comes to encoding data on a computer, it is important not to let one particular view unduly influence the way the data itself is coded. Each view is optimized for a particular use and encoding some data too closely to one particular view on a computer will make it hard for it to be reused to create other views. Instead, one should attempt an analysis of the underlying logical structure of the data being collected, encode it using that logical structure, and then allow existing software tools to create views of the data of use to the various interested communities and individuals.

Section 4 will cover specific issues relating to the encoding of language data on a computer. In the remainder of this section notion of an underlying data structure will be explored in more detail (section 3.1) and general aspects of the

[9] We should clearly distinguish here between encoding data in non-proprietary, open formats which, in principle, allow it to be straightforwardly repurposed by outside parties and actually making the data available to them, for example by posting it on a website. Open access and open formats are distinct concepts, and neither implies the other.

problem of encoding that structure in machine-readable format will be introduced (section 3.2). For purposes of illustration, the discussion will focus on the structure of a simple entry in a wordlist.

3.1 Underlying data structures

In trying to determine what the basic underlying structure is for a given kind of data, the first point one must keep in mind is that this is a complex analytical task and developing a universal mechanistic algorithm to determine the underlying structure of language data is no easier than, say, developing such an algorithm for discovering the phoneme inventory of a language based on phonetic transcriptions. Each kind of data from each language will present its own conceptual difficulties, though just as with grammatical analysis, these will often be variations on a theme rather than completely unexplored problems.

To make the discussion more concrete, consider the very simple lexical entry in (1), associating a French word with a part of speech and an English translation. (See Austin 2006:97–98 for comparable discussion of the structure of a lexical entry.)

(1) *chat* **n.** cat

The example in (1) gives a particular view of a bilingual lexical entry consisting of a headword from the language being described in italics, an indication of its part of speech in bold, and a basic translation in plain text. The underlying structure of the data is largely implicit, though the view does at least imply that the data can be analyzed into three core pieces. We can give a first approximation of the underlying structure of the data in (1) as in (2), where the typological conventions of (1) are repeated in the interests of clarity.

(2) *headword* **pos** gloss

While (2), at first, may seem to be a reasonable representation of the logical structure of (1), it, in fact, still leaves many characteristics of the data itself implicit. This is because it only analyzes those features of the data explicitly represented in the view seen in (1), leaving out many important other features, which, while easily reconstructible from context by a human, will be unknown to a computer without explicit coding. Perhaps the most important of these implicit features is the most easily overlooked: the three logical pieces in (2) are part of a larger unit we might refer to as an *entry*, and represent as in (3).

(3) [[*headword*] [**pos**] [gloss]]$_{\text{ENTRY}}$

There is at least one set of important additional characteristics associated with the entry in (1) not yet described by the analysis in (3)—that each of the parts of the entry is associated with a particular language. The headword is in French, the part of speech label is an abbreviation from English (though an abbreviation like *n* is, of course, potentially ambiguous as to what language it is drawn from), and the gloss is in English. We might, therefore, want to expand our analysis of the underlying structure of the word list entry in (1) as in (4).

(4) [[*headword*]$_{\text{lang:french}}$ [**pos**]$_{\text{lang:english}}$ [gloss] $_{\text{lang:english}}$]$_{\text{ENTRY}}$

While (4) is significantly more complex than (2), it is still just a beginning. Nowhere is it explicitly indicated yet, for example, that that part of speech label applies to the headword and not to the gloss. Nor is there indication of the nature of the representation of the headword—that is, we do not know (without using outside knowledge) whether the sequence *chat* is a phonetic, phonemic, or orthographic representation.

Should we further refine the analysis given in (4), then? How one answers this depends on the details of the data being collected as well as what the data will be used for. For example, if one was working with a dataset wherein some of the headwords were given in an orthographic representation while others were given in phonetic transcription, then it would be important to include the possibility for specifying the nature of the headword's representation in an analysis of the entry's underlying structure. However, if all the headwords used an orthographic representation, this would be relatively less important.

3.2 Implementing a data structure

Analyzing some data in order to arrive at an understanding of its underlying structure could, in principle, be a purely theoretical enterprise. However, in language documentation, it is mostly a means to an end: What one wants to be able to do is store data on a computer in a form which will facilitate its being used to produce human-usable language resources. Therefore, there will generally be a point when some analysis of this structure, even one that may be known to be imperfect, must be chosen for *implementation* on a computer—that is, a method must be devised for it to be expressed in a machine-readable form which can be straightforwardly manipulated by the user.

Deciding on an implementation for a given data structure, ultimately, is largely dependent on practical considerations relating to the intended uses for the data and the range of data manipulation tools available to the language documenter. Nevertheless, it is still essential to devote some time to abstract data modeling of the sort described in section 3.1. Simply put, the better one understands the underlying structure of one's data, the easier it will be to arrive at an implementation which will be sustainable over the lifespan a given project.

An implementation of a data structure by definition will need to be done using some computational tool. From the present perspective, one of the most crucial factors in choosing a tool is that it will be able to straightforwardly create a reasonable implementation of the underlying data structure one chooses to work with. In that sense, one of the most ubiquitous kinds of application, the word processor, is usually insufficient since word processors are optimized to work with a kind of data—unannotated text documents—that plays a relatively minor role in language documentation. Thus, while one may be able to create reasonable presentations of data (see section 4.3), like what is seen in (1) using a word processor, the resulting resource will not actually code the structure of the data but, rather, aspects of formatting (e.g., bold and italics) that are only indirectly related to the structure. Another common office application, spreadsheet software, by contrast, can be used profitably to implement data structures which are well expressed in a table. The crucial issue here is not the fact that each of these products was designed for use in an office environment. Rather, it is that one kind of application (spreadsheet software) builds a basic kind of data structure (the table) directly into its design.

Software specifically designed for language documentation will be optimized to work with a particular linguistic data type (or set of data types)—e.g., time-aligned annotated texts in the case of Elan. But, such software will not be available for every kind of data and, depending on the needs of a project, may not always be the ideal choice, particularly when a documentary team consists of not only linguists but also non-linguists, who might not be familiar with the ways that linguists think about their data which inform the design of the linguistics-specific tools.

Returning to the example of a lexical entry discussed in section 3.1, how might we implement the data structure associated with it? In this case, the structure is relatively simple, and we could straightforwardly implement it in a spreadsheet where each row corresponds to an entry, and where each part of the entry occupies a single cell of the row, along the lines of what is depicted in table 1. (See section 4.2 for an alternative way of encoding the data.)

Table 1

Tabular implementation of word list entries

headword	**part of speech**	**gloss**
chat	n.	cat
chien	n.	dog

The implementation in table 1 does not contain all the information found in the underlying data analysis presented in section 3.1. For example, there is no specific indication that the headword is French and the glossing language is English. Some of the structure is explicitly indicated, however, in the header line which labels the uses of each column. In this case, the missing language information does not pose particular problems since it could be straightforwardly rectified with accompanying information documenting the nature of the data in the file, which could be as easy as giving the spreadsheet a title like "French wordlist with English glosses". In this case, we are dealing with data that has a relatively simple structure and which, therefore, can be a given a fairly simple implementation using a widely available kind of software.

Of course, this is just an illustrative example. In many—perhaps most—cases the data collected while documenting a language will be more complex than the example given in (1). Bell and Bird's (2000) survey, for example, of the structure of lexical entries across a wide range of published work gives a good indication of the level of complexity involved when one looks at real lexical data. A full dictionary entry—as opposed to word list entry—which might contain multiple senses of a given word, example sentences for each sense, and comparative notes, among other things, will require a tool allowing the definition of data structures with hierarchical relationships within an entry, for example linguistics-specific database software like SIL International's Shoebox/Toolbox or commercial database software like FileMaker Pro. Similarly, in a language documentation project, one will often want to create machine-readable representations of the relationship between textual data and audio or video recordings (e.g., in the form of time-aligned transcription). Doing this requires software which allows one to make direct associations between portions of distinct computer files—something beyond the power of a spreadsheet program but which is made easy with a tool like Elan.

While the use of linguistics-specific software will generally facilitate the creation of implementations that are faithful to the underlying structure of the data, simply using such software does not guarantee that the data will come out “right”. For instance, a lexicon tool may make it straightforward to specify morphosyntactic information like part of speech, but in a language where it is deemed valuable to list multiple paradigmatic forms of a word within a lexical entry, one may want to indicate not only a part of speech at the level of the lexeme but also associate each word form with additional grammatical categories (e.g., a case label). This requires a two-tiered model of grammatical specification, at the lexeme level. A given lexicon creation program may support this, but it cannot “know” to make use of such a feature unless the documenter is aware that it is needed in the first place. A “perfect” implementation of a flawed analysis of the structure of some data will be of little long-term value and, at least for now, arriving at good structural analyses of linguistic data is a task well beyond the skills of any machine.

It would be ideal, of course, if, in a chapter like this one, it would be possible to give explicit recommendations about what software is “best” for language data of a particular type. Unfortunately, the needs of every project are too particular for this to be possible, and there is a tradeoff between being able to implement a data model as faithfully as possible to its underlying logical structure, employing a tool that everyone on a project team can use comfortably, and ensuring that the tool that is used can produce resources which can be put to use by the audiences to be served by a project. The main advice one can give is to outline the overall goals of a project and data types to be collected in advance (see section 6) and then to solicit advice from experienced individuals when making choices of software. One important factor to consider when choosing software will be the kinds of formats it is able to work with (see section 4).

3.3 Audio and video resources and publications

It may seem like a gap in the discussion in this section that it has focused on “traditional” text-oriented resources rather than recordings. There is a reason for this: Many of the important components of the documentary record of a language employ data types which are of interest to communities well outside of the arena of language documentation and which, therefore, will be well-supported independent of language documentation efforts. Audio and video recordings are a prime example of this: Technologies for capturing, storing, and manipulating audio and video data have a large, stable market of which language documentation work is only a minute part. Therefore, efforts will be made to model the structure audiovisual information and implement those models regardless of the activities of language documenters.

Publication technologies are similar in this regard. The audience for old (e.g., print publications) and new (e.g., multimedia content) modes of information dissemination is vast and new models and technologies for producing publications—in a broad sense of the term—will emerge with or without language documentation work. Therefore, given limited resources, language documenters will need to devote more energy to issues relating to the modeling and implementation of data types specific to documenting languages, like annotated texts, lexicons, and grammars. Nordhoff’s (2008) discussion of a possible set of design principles and implementation decisions for the creation of “ideal” electronic grammars is a good recent example of the kind of work which is needed.

4. DATA FORMATS

Closely related to the notion of data model implementation is the notion of data *format*, that is, the way that information happens to be encoded in a digital resource. When using this term, we must first recognize that it is potentially quite vague and is better understood as a multidimensional concept referring to a number of distinct “layers” of data encoding rather than a single monolithic notion. In particular, in the present context it is useful to distinguish between *file format* and *markup format*. The former concept is likely the more familiar since it refers to the different file types associated with software applications. These include, for example, the DOC format created by Microsoft Word, PDF format, or WAV audio format. The details of the structure and digital composition of these formats are largely irrelevant to language documenters, though, as will be discussed in section 4.1, some are more suitable for language documentation than others. By contrast, markup format, in the present context, refers to the way the substantive content (at least from the documenter’s perspective) of a resource is encoded on top of a particular file format. As such, it is directly relevant to language documenters and will be discussed in more detail in section 4.2. In section 4.3, a third way of categorizing formats, by their intended function, will be discussed.

This section will focus primarily on conceptual issues relating to data formats. For specific recommendations regarding appropriate formats to use for different kinds of data (e.g., text, audio, or video) and for different kinds of functions (e.g., archiving versus presentation), it is best to refer to up-to-date online resources (e.g., the E-MELD School) or to contact a digital archivist or other individual with the relevant expertise. Standards recommendation for digital formats tend to evolve rapidly, and periodic review of the state-of-the-art is required for successful language documentation. Video formats, in particular, have yet to see the same degree of stabilization as text and audio formats.

4.1 File formats: Open versus proprietary

The most important way in which file formats can differ from the perspective of language documentation is whether or not they are *open* or *proprietary*. Devising satisfactory definitions of these terms is not completely straightforward, but, practically speaking, the distinction centers around whether a given format is designed to be used in any application which may find that format a useful way to store data or whether it is intended to be used only by the format’s owner or via licensing agreements with that owner.

Among the most widely-used open file formats is the “raw” text file (sometimes referred to as a TXT file or by the file extension .txt), consisting of a sequence of unformatted characters—these days, ideally, of Unicode characters (see Anderson 2003 and Gippert 2006:337–361 for an overview of Unicode). Such files can be created and read by a wide array of programs on all widely used operating systems, and no one organization has any kind of ownership over the format. By contrast, a well-known proprietary format is the Microsoft DOC format. While this format is creatable and readable by programs not created by

Microsoft, it was not designed specifically for this, and the format has been subject to change under Microsoft's discretion regardless of how this may have impacted the ability for other software to create and read files in that format.[10]

For work on language documentation, one of the most important recommendations is to prefer the use open formats whenever possible, and always for the archival version of a resource (see section 4.3). There are two major reasons for this. First, open formats, by their nature, are more likely to be created and read by different computer programs, which means that resources encoded in open formats will generally be available to a wider audience than proprietary formats. Furthermore, open formats are much more likely to be supported by cost-free programs since, very often, the reason why a format is proprietary in the first place is so a company can profit from selling software which can work with files in that format. While the issue of cost may not be particularly relevant to linguists working at well-funded universities, one must keep in mind that the larger audience for a documentary resource will often consist of individuals or groups which are not particularly privileged financially.

The second reason to disprefer proprietary formats is that, by virtue of being largely under the control of a particular company, they are more likely to become obsolete—that is, resources encoded using them are more likely to become unreadable or uneditable—because the company controlling them may decide to change the format that its tools support over time, while discontinuing support for its earlier formats, or because the company itself may disappear, meaning that its formats will no longer be supported by any program. With open formats, even if one institution making a tool supporting that format should cease to exist, the nature of the format itself makes it relatively easy for a new group to create a tool supporting use of that format.[11]

4.2 Markup formats

Markup, in a digital context, refers to the means by which part of the content of a given document is explicitly "marked" as representing some type of information. Continuing the example of a wordlist entry discussed in section 3.1, markup could be used to indicate, among other things, that: (i) the data in question is a lexical entry, (ii) the first element of the lexical entry is the headword, (iii) the second element is an indication of part of speech, and (iv) the third element is a gloss.

An example of the data in (1) presented in a possible markup format is given in (5), where a markup language known as Extensible Markup Language (XML) is used. XML is a widely used open standard for marking up data using a system of start and end tags which surround data of the type specified by the tag. The distinction between a start and an end tag is maintained by the prefixation of a slash before the name of an end tag. Start tags can have complex structure wherein they include not only the tag but also specification of attributes of the

[10] In recent years, the DOC format has been replaced by the DOCX format which, in principle, is an open file format—though, in practice, it has not yet been widely adopted outside of Microsoft.

[11] It is important to distinguish between open source and open format. Open source refers to whether or not the computer code that forms the basis of a program is made freely available for inspection and modification. In practice, open source programs are more likely to use open formats for various reasons, some practical and some social. However, many closed-source programs also allow one to produce resources in open formats (e.g., Microsoft Word allows one to save documents into the open HTML format)

data using feature-value pairs indicated with equal signs. In (5) these are used to specify the language of the content of the tags. Readers familiar with HyperText Markup Language (HTML), the dominant markup format for web pages, should find the overall syntax of XML to be familiar since the two use the same basic conventions (see Gippert 2006:352–361 for additional relevant discussion).

(5)

```
<lexicalEntry>
        <headword lang=“French”>
        chat
        </headword>
        <pos>
        n.
        </pos>
        <gloss lang=“English”>
        cat
        </gloss>
</lexicalEntry>
```

The XML in (5) is somewhat simplified for purposes of exposition. Nevertheless, it gives a basic idea of data markup in general and XML specifically. While numerous markup languages have been developed, XML has been chosen here for illustration since, at present, it enjoys widespread popularity within the software development world as a format facilitating the exchange of data across individuals and computer programs and is considered an appropriate markup format for language data where markup is relevant.

XML has at least four attributes which make it especially well suited for language documentation. First, it can be expressed in plain text—i.e., the markup tags do not use any special characters or formatting not found in plain text files. This means that XML files can make use of a widely-adopted open format and facilitates archiving. Second, while XML is primarily designed to be a machine-readable markup format, the fact that the tags can make use of mnemonic text strings (e.g., “lexicalEntry” in (5)) means that it can be, secondarily, human-readable. Thus, even in the absence of materials documenting the specific markup conventions used in a given resource, it will still often be possible to discern the content of a document marked up with XML by inspecting it with a simple text editor. This self-documenting feature of XML markup is a desirable characteristic for the long-term preservation of the data in the document since it helps ensure its interpretability even if a document becomes detached from its metadata (see section 5). Third, XML is flexible enough to mark up a wide range of data types for diverse kinds of content—one simply needs to define a new kind of tag to mark up a new kind of data. Finally, XML has been widely-adopted in both commercial and non-commercial contexts. As a result, there is extensive tool support for processing and manipulating XML documents, going well beyond what would be possible to create with the resources solely devoted to language documentation.

While the XML example in (5) may make it appear to be a markup format of use only to specify the data contained in resources which would traditionally be printed (e.g., dictionaries or texts), it can also be used to annotate other kinds of resources, like audio and video recordings or images using so-called *stand-off* markup, wherein the markup itself is stored in a separate resource from the resource it describes. Such stand-off markup can then specify which part of an external resource it refers to using some kind of "pointer", for example the specification of horizontal and vertical coordinates in a scanned image. A common use of such stand-off markup in language documentation is to create a time-aligned transcription of a recorded text where the text transcription is encoded in an XML file containing pointers to times in an audio file—as is done in the EAF files produced by the Elan annotation tool (while these files end in the extension .eaf rather than .xml, the data contained within them is expressed in XML).

While use of a markup language like XML solves many problems associated with describing the content of a language resource, it is important to understand that, on its own, it is merely a scheme for marking data with different kinds of tags—not, for example, a standardized way of encoding lexical data or an annotated text. Rather, one must, beforehand, develop an abstract model of a lexicon or a text, and then implement it in XML (see section 3 for discussion of modeling and implementation). XML—or any generalized markup language—serves merely as a kind of "skeleton" on which domain-specific markup schemes can be constructed. In the long run, the creation of long-lasting, repurposable language documentation will be greatly facilitated by the use of common markup conventions for basic linguistic data types, which will allow for the development of tools which can work with the data from diverse documentation projects making use of these conventions. At present, however, general consensus has yet to emerge for most aspects of the markup of linguistic data.[12] In the absence of such consensus, the best strategy is to employ markup conventions using mnemonic labels and to document how those labels are to be interpreted in the context of a given resource.

Finally, in general, one will not manipulate markup directly, for example by editing an XML document in a text editor. Rather, one will use software providing a graphical interface to the markup (as Elan does with its XML format, for example) or software which allows for the data it creates to be exported to an appropriate markup format—as is the case with, for example, FileMaker Pro's XML export. However, while one need not learn how to create or edit a suitable markup format directly, it is important to be able to determine whether a markup format is sufficiently open and transparent to be appropriate for a project's documentary needs, which requires some knowledge of the relevant issues.

4.3 Archival, working, and presentation formats

In addition to classifying formats by their various technical features, one can also classify a format by virtue of its possible or optimal functions. In the context of language documentation, three particular functions stand out: *archival*, *working*, and *presentation*. An archival format is one designed for longevity. In the ideal case, a resource stored in an archival format today would be readable in a hundred years or more (assuming it has not been lost on unreadable media). A working format is one manipulated by a given tool as the user creates or edits a resource—this is the format language documenters will spend most of their time with. A presentation format is a version of the resource optimized for use by a specific community. Presentation formats can range from a print dictionary to a

[12] To take one example, despite being fairly well-studied, consensus has yet to emerge on the ideal markup format for interlinear glossed text (see Palmer and Erk (2007) for recent discussion).

multimedia text presentation and are what those not involved in the language documentation process itself would generally consider to be the "normal" kind of language resource. For discussion of archival, working, and presentation formats for different data types referencing specific formats, consult the E-MELD School.

In an ideal world, a single format could function simultaneously as an archival, working, and presentation format for a given kind of resource. However, this is a practical impossibility. This is most clearly the case for presentation formats which are, by definition, audience specific (e.g., an ideal linguist's dictionary has a very different form from a community dictionary, even if they can be based on the same underlying lexical database) and also may require optimization for certain modes of dissemination (e.g., an audio file may need to be reduced in size, and therefore quality, in order to become suitable for distribution via the internet). Though such problems are not as acute when comparing archival formats and working formats, they do not disappear entirely. For example, archival formats often tend to be large and "verbose"—that is, they may express their content with lots of redundancy—since this helps ensure their long-term readability. Working formats, by contrast, are often more useful if expressed in ways that are concise, since this allows them to be manipulated more efficiently by a computer.

A language documentation project, therefore, needs to anticipate the use of formats with distinct functions over its lifespan, working formats for performing day-to-day tasks, archival formats for long-term storage, and a variety of presentation formats depending on the communities it wishes to serve and the ways it wishes to serve them. The need for such a variety will inevitably complicate the management of a documentation project, though such complications can be alleviated by forward planning (see section 6) and the use of tools either natively using open formats as working formats or allowing easy and reliable export of their working format to an open format since such formats tend to be more straightforwardly transformable to appropriate archival and presentation formats than proprietary formats.

5. METADATA

In order for the data collected by a project to be usable in the long-term, it not only needs to be well-structured internally but also must be associated with appropriate *metadata*—that is, information describing the constituent resources of a documentary corpus, including, for example, their content, creators, and access restrictions (see Good (2003) for introductory discussion in a linguistic context). Metadata is an essential part of any documentary corpus, and a metadata plan forms an integral part of a general data plan.

Since materials deposited in an archive will need to be associated with their metadata in order for them to be accessioned into an archive (see Conathan (this volume)), the best place to turn to for advice in terms of what metadata you should include with your resources is the archive where you will deposit your data, assuming it is clear what archive is best placed to protect the resources created by your project. While the metadata policies for language archives are all broadly similar, each archive will have its own specific expectations and, in some cases, an existing set of forms which can be used for metadata entry and which the archive will design to facilitate its own accessioning process.

In devising a metadata plan for a language documentation project, it is useful to think about your metadata needs across two broad parameters: the different kinds of items that will require metadata and the different users of your metadata. I will not consider here in detail the specific metadata "fields" one may want to record, since there are a number of complicated considerations involved relating to specific project requirements and resources (though see Conathan (this volume, section 3.2) for relevant suggestions). At a minimum, it is necessary to record basic "bibliographic" information like creators (a cover term encompassing anyone involved in a resource's creation), date of creation, place of creation, language being documented, access restrictions, and brief descriptive title or keyword (see Johnson 2004:250). At a maximum, one can consider the extensive IMDI[14] metadata set—most projects will fall somewhere in between. If you are starting a new project, it may be useful to look at the latest version of the IMDI set to get an idea for the range of information that, in principle, might be worth keeping track of.

5.1 What requires metadata

Most of the documentary objects requiring metadata can be arranged in a hierarchy from more general to more specific using the categories *project*, *corpus*, *session*, and *resource*.[15] An additional set of "objects" requiring metadata, but which do not fit directly into this hierarchy, are the various *people* involved, including most prominently speakers and documenters.

A *resource*, in this context, is a unique object, either a physical item or a computer file, comprising part of the documentation of a language. Often multiple resources are created as part of the record of a single event (e.g., an audio recording, a transcription, and an associated photograph). These would then be grouped into a *session* (following the terminology adopted by IMDI as discussed in Brugman et al. (2003), though the term *bundle* is also used for this concept). Sessions may then belong to some user-defined higher-level grouping which can be referred to as a *corpus*, which might, for example, consist of all sessions documenting a specific language in a multilingual documentation project. Finally, a set of corpora may be joined together into a larger *project*, for example all the materials collected by a given documentary team. While it is generally possible to apply the notions *resource* and *session* fairly consistently, *corpus* and *project* are somewhat more subjective and are more likely to be employed using conventions specific to a documentary team.

Conceiving of the items produced by a language documentation project as belonging to a hierarchy is useful insofar as it allows one to avoid repeating the same information in multiple places. For example, if documentary work is externally funded, it will often be necessary to acknowledge that funder somewhere in the metadata. This is most conveniently done at a high-level, like that of *project*, as opposed to specifying this for each individual resource. Similarly, resources documenting a single speech event will share information like *creators* and *date*, thus making it useful to employ the notion of *session*. Finally, since most information about people is independent of the actual resources they contributed to, person metadata constitutes a level on its own. Each

[14] http://www.mpi.nl/imdi/

[15] The conceptual metadata scheme discussed here is derived from work done in the context of IMDI. See Brugman et al. (2003).

person can be associated with a unique identifier (e.g., their name, if appropriate), which can then be referred to in session metadata.

5.2 Metadata users

When creating metadata, one should consider the range of users who are likely to make use of it, with the most important division being those directly involved in a project versus those outside of it. On the one hand, those involved in a project are unlikely to be, for example, interested in project-level metadata since they will already be aware of such information. By contrast, they are likely to be very interested in session-level metadata as a means to keep track of a project's progress. On the other hand, those outside of a project are likely to want to refer to project-level metadata as a first "entry point" into a set of documentary materials and will only be interested in session-level metadata for projects which they have determined are relevant to their interests.

A documentary team will presumably keep track of the metadata it needs for its own purposes without special consideration but may forget to record information that is shared among the team but will be unknown to outsiders. For example, the fact that a given speaker is an elder will be obvious to those working directly with that speaker but could be very difficult to determine from an audio recording. Therefore, the language documenter must try to keep in mind that the users of metadata are not privy to the same level of information that a documentary team will be. In fact, the concerns of one particular group of "outside" users should resonate particularly strongly with experienced documenters: Future versions of themselves who are likely to forget quite a bit about the context of their old recordings but will still be interested in using them.

This two-way distinction between project members and those outside of a project is, of course, quite simplistic and masks many internal divisions within those categories. With respect to outsiders, a further important division involves researchers versus community members. Existing metadata schemes for language resources, like IMDI (see above) and the Open Language Archives Community metadata set (OLAC; Simons and Bird (2008)) are oriented towards the research community, and speaker communities are likely to have distinct interests in terms of the information they find valuable. For example, linguists are typically more concerned with the languages a given speaker's parents may have spoken at home than they are with who that person's parents actually are, while speaker communities are quite likely to be interested in the genealogical relations of those who participated in the creation of a set of documentary resources—especially if they are close relations.

5.3 Practical considerations

While it is not possible here to go into details regarding metadata management techniques, two practical considerations are especially crucial. First, every resource created by a documentation project should be associated with a unique identifier. For computer files, this identifier should be the name of the file itself, which, therefore, needs to be created with uniqueness in mind. For physical resources, this identifier should be marked on the resource itself directly or with an adhesive label. (See Johnson 2004:149–151 for examples of possible schemes for creating unique identifiers relevant to a language documentation context.) In an ideal world, a given resource would be indelibly associated with its metadata so that its content would always be completely clear. However, in practice, metadata tends to be stored separately from the resource itself. Therefore, it is also useful for a resource's identifier to give some minimal information about its content. Then, even if the resource cannot be straightforwardly associated with its metadata at a given time, some information about it can be gleaned from its label. For example, a recording of Angela Merkel in German made on 1 January 2009 might have a label like *deu-AM-20090101.wav*. This identifier contains a three-letter language code, followed by the initials of the speaker, a date, and, finally, a file extension indicating this is a WAV audio recording. Obviously, such an identifier does not substitute for a full metadata record, but it, at least, gives some information about a resource which will be quite valuable in case its metadata becomes lost.[16]

A second practical consideration regarding metadata is that, especially in field settings, it is essential that metadata entry be made as straightforward as possible. Ideally, metadata will be recorded for a resource on the same day it is created—while one's memory is still fresh. But, language documentation can often be a tiring task, leaving little energy at the end of the day to work with a complex metadata management system. Since metadata usually has a fairly simple structure almost any program one might use to create a table or a database, e.g., Microsoft Excel, FileMaker, or Shoebox/Toolbox, can be used for metadata entry and storage. Since one such tool is already likely be used for other aspects of documentation, the most straightforward route is to co-opt it for use as a metadata entry and storage tool as well—at least when in the field.[17]

6. NEEDS ASSESSMENT

Implicit in the discussion to this point has been that, either formally or informally, a given project has undertaken a technical needs assessment—that is, the overall goals of a project have been outlined, an enumeration of the different resources required to reach those goals has been formulated, and a workplan has been devised to ensure that those resources can be acquired or developed over the course of the project. Bowern (this volume) contains a general overview of issues relating to project planning, including some discussion of how to integrate a project's data needs into its overall design.

A useful notion to keep in mind while considering the data management aspect of a needs assessment is the *workflow* of the individuals involved in the project: That is, what will be the series of day-to-day tasks each project participant will work on at each phase of the project. Modeling a project's workflows will help ensure that the optimal technologies are chosen to accomplish its goals since it will clarify the specific technological needs of each member of the project team. So-called "lone wolf" research may only require an informal understanding of a project's workflow, while projects involving large and diverse teams may benefit from a more formalized depiction of workflow breaking down project work into a set of interconnected tasks. A very large project may even require a member of the documentary team to invest substantial (paid) time in managing its overall workflow.

[16] For similar reasons, it is often helpful to record some brief metadata at the beginning of an audio or video recording.

[17] The Archive of Indigenous Languages of Latin America (AILLA) has examples of Excel spreadsheets and Shoebox/Templates which can be used for metadata management.

7. THE DOCUMENTER'S RESPONSIBILITY

This chapter can only give a brief outlines of the relationship between data and language documentation. Furthermore, because the technologies for capturing and storing data are continually evolving, our understanding of data in the context of language documentation will also continually evolve, and the language documenter will have to periodically reconsider their technological practices and keep abreast of new developments by consulting up-to-date resources.

Unlike, say, learning how to transcribe using the IPA, working with the data produced by language documentation is not something you can simply "learn once". Rather, it will be an ongoing, career-long process. Furthermore, since, in many cases, the access that many individuals leading language documentation projects have to new technologies greatly exceeds that of the communities they work with, it is, to some extent, their responsibility to serve as the conduit through which information about these technologies reaches these communities (see Jukes (this volume) for relevant discussion).

The most succinct way to summarize these points is: understanding how data collection and management fits into a documentation project is a kind of *research*. It, therefore, submits to all the requirements of research: keeping up with the field, knowing the limits of one's expertise, tracking down outside sources, constantly evaluating and reevaluating one's conceptual understanding and methodological practices, and instructing collaborators on appropriate practices. Just as analyzing your data requires research, so does working with the data itself.

REFERENCES

Anderson, Deborah. 2003. Using the Unicode standard for linguistic data: Preliminary guidelines. *Proceedings of the E-MELD Workshop 2003: Digitizing and Annotating Texts and Field Recordings*, 10pp. <http://www.emeld.org/workshop/2003/anderson-paper.pdf>

Austin, Peter K. 2006. Data and language documentation. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 87–112. Berlin: Mouton de Gruyter.

Bell, John & Steven Bird. 2000. A preliminary study of the structure of lexicon entries. *Proceedings from the Workshop on Web-Based Language Documentation and Description*. Philadelphia, December 12–15, 2000. <http://www.ldc.upenn.edu/exploration/expl2000/papers/bell/bell.html>

Berez, Andrea. 2007. Technology review: EUDICO Linguistic Annotator (ELAN). *Language Documentation and Conservation* 1:283–289. <http://hdl.handle.net/10125/1718>

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79:557–582,

Boas, Franz (ed.). 1911. *Handbook of American Indian languages*. Washington: Government Printing Office. (Smithsonian Institution Bureau of American Ethnology Bulletin 40.)

Boynton, Jessica, Steven Moran, Anthony Aristar & Helen Aristar-Dry. 2006. E-MELD and the School of Best Practices: An ongoing community effort. In:

Linda Barwick and Nicholas Thieberger (eds.), *Sustainable data from digital sources: From creation to archive and back*. Sydney: Sydney University Press. <http://hdl.handle.net/2123/1296>

Brugman, Hennie, Daan Broeder & Gunter Senft. 2003. Documentation of Languages and Archiving of Language Data at the Max Planck Institute for Psycholinguistics in Nijmegen. Paper presented at the "Ringvorlesung Bedrohte Sprachen" Sprachenwert - Dokumentation - Revitalisierung, Fakultät für Linguistik und Literaturwissenschaft - Universität Bielefeld, 17pp. <http://www.mpi.nl/IMDI/documents/articles/BI-EL-PaperA2.pdf>

Dryer, Matthew. 2006. Descriptive theories, explanatory theories, and basic linguistic theory. In Felix Ameka, Alan Dench & Nicholas Evans (eds.), *Catching language: The standing challenge of grammar writing*, 207–234. Berlin: Mouton de Gruyter.

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 31–66. Berlin: Mouton de Gruyter.

Good, Jeff. 2003. A gentle introduction to metadata. Open Language Archives Community Note. <http://www.language-archives.org/documents/gentle-intro.html>

Grinevald, Colette. 2003. Speakers and documentation of endangered languages. In Peter Austin (ed.), *Language documentation and description, volume 1*, 52–72. London: Hans Rausing Endangered Languages Project.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36:161–195.

Himmelmann, Nikolaus P. 2006. The challenges of segmenting spoken language. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 253–274. Berlin: Mouton de Gruyter.

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. In Peter Austin (ed.), *Language documentation and description, volume 2*, 140–143. London: Hans Rausing Endangered Languages Project.

Mithun, Marianne. 2001. Who shapes the record: The speaker and the linguist. In Paul Newman & Martha Ratliff (eds.), *Linguistic fieldwork*. Cambridge: CUP.

Mosel, Ulrike. 2006. Sketch grammar. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 301–309. Berlin: Mouton de Gruyter.

Nathan, David. 2006. Thick interfaces: Mobilizing language documentation with multimedia. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 363–379. Berlin: Mouton de Gruyter.

Nordhoff, Sebastian. Electronic reference grammars for typology: Challenges and solutions. *Language Documentation and Conservation* 2:296–324. <http://hdl.handle.net/10125/4352>

Palmer, Alexis and Katrin Erk. 2007. IGT-XML: An XML format for interlinearized glossed texts. *Proceedings of the Linguistic Annotation Workshop*. Prague: Association for Computational Linguistics. 176–183. <http://www.aclweb.org/anthology/W/W07/W07-1528>

Schultze-Berndt, Eva. 2006. Linguistic Annotation. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 213–251. Berlin: Mouton de Gruyter.

Simons, Gary & Steven Bird (eds.). 2008. OLAC Metadata. Open Language Archives Community Standard. <http://www.language-archives.org/OLAC/metadata.html>

Simons, Gary, Kenneth S. Olson & Paul S. Frank. 2007. Ngbugu digital wordlist: A test case for best practices in archiving and presenting language documentation. *Linguistic Discovery* 5:28–39. <http://journals.dartmouth.edu/cgi-bin/WebObjects/Journals.woa/2/xmlpage/1/article/314>

Seven Dimensions of Portability for Language Documentation and Description

Steven Bird, Gary Simons

LANGUAGE
A JOURNAL OF THE LINGUISTIC SOCIETY OF AMERICA

Language, Volume 79, Number 3, September 2003, pp. 557-582 (Article)

Published by Linguistic Society of America
DOI: 10.1353/lan.2003.0149

SEVEN DIMENSIONS OF PORTABILITY FOR LANGUAGE DOCUMENTATION AND DESCRIPTION

Steven Bird
University of Pennsylvania and University of Melbourne

Gary Simons
SIL International

The process of documenting and describing the world's languages is undergoing radical transformation with the rapid uptake of new digital technologies for capture, storage, annotation, and dissemination. While these technologies greatly enhance our ability to create digital data, their uncritical adoption has compromised our ability to preserve this data. Consequently, new digital language resources of all kinds—lexicons, interlinear texts, grammars, language maps, field notes, recordings—are proving difficult to reuse and less portable than the conventional printed resources they replace. This article is concerned with the portability of digital language resources, specifically with their ability to transcend computer environments, scholarly communities, domains of application, and the passage of time. We review existing software tools and digital technologies for language documentation and description, and analyze portability problems in the seven areas of content, format, discovery, access, citation, preservation, and rights. We articulate the values that underlie our intuitions about good and bad practices, and lay out an extensive set of recommendations to serve as a starting point for the community-wide discussion that we envisage.*

1. Introduction. Language documentation provides a record of the linguistic practices of a speech community, such as a collection of recorded and transcribed texts. Language description, on the other hand, presents a systematic account of the observed practices in terms of linguistic generalizations and abstractions, such as in a grammar or analytical lexicon.[1] It is now easy to collect vast quantities of language documentation and description and store them in digital form. It is easy to transcribe the material using appropriate scripts, to organize it into databases, and to link it to linguistic descriptions. It is also easy to disseminate rich language resources on the internet. Yet how can we ensure that this digital language documentation and description can be reused by others, both now and in the future?

Today's linguists can access printed and handwritten documentation that is hundreds (sometimes thousands) of years old. However, much digital language documentation and description becomes inaccessible within a decade of its creation. Linguists who have been quick to embrace new technologies, create digital materials, and publish them on the web soon find themselves in technological quicksand. Funded documentation projects are usually tied to software versions, file formats, and system configurations having a lifespan of three to five years. Once this infrastructure is no longer tended, the language documentation is quickly mired in obsolete technology. The issue is acute for endangered languages. In the very generation when the rate of language death is at its peak, we have chosen to use moribund technologies, and to create endangered data. When the technologies die, unique heritage is either lost or encrypted. Fortunately, linguists can follow best practices in digital language documentation and description, greatly increasing the likelihood that their work will survive in the long term.

* This research was supported by NSF Grant No. 9983258 'Linguistic Exploration' and Grant No. 9910603 'International Standards in Language Engineering (ISLE)'. We are grateful to Dafydd Gibbon, David Nathan, Nicholas Ostler, and the *Language* editors and anonymous referees for comments on earlier versions of this article.

[1] For a lucid discussion of the terms 'language documentation' and 'language description' we refer the reader to Himmelmann 1998.

If digital language documentation and description should transcend time, they should also be reusable in other respects: across different software and hardware platforms, across different scholarly communities (e.g. field linguistics, language pedagogy, language technology), and across different purposes (e.g. research, teaching, development). In this article we address all these facets of the problem under the heading of PORTABILITY. Portability is usually viewed as an issue for software, but here our focus is on data. By 'data' we mean any information that documents or describes a language, such as a published monograph, a computer data file, or even a shoebox full of handwritten index cards. The information could range in content from unanalyzed sound recordings to fully transcribed and annotated texts to a complete descriptive grammar.

This article addresses seven dimensions of portability for digital language documentation and description, identifying problems, establishing core values, and proposing best practices. The article begins with a survey of existing tools and technologies, leading to a discussion of the problems that arise with the resources created using these tools and technologies. We identify seven kinds of portability problems under the headings of CONTENT, FORMAT, DISCOVERY, ACCESS, CITATION, PRESERVATION, and RIGHTS. Next we give statements about core values in digital language documentation and description, leading to a series of VALUE STATEMENTS that serve as requirements for best practices. Finally, we discuss OLAC, the Open Language Archives Community, which provides a process for identifying community-agreed best practices, and lay out an extensive set of recommendations to serve as a starting point for the community-based effort that we envision.

The structure of the article is designed to build consensus. Readers who take issue with a best practice recommendation in §6 are encouraged to review the corresponding statement of values in §4 and either suggest a different practice that better implements the values, or else propose a more appropriate value statement. The reader could turn further back to the corresponding problem statement in §3 and offer a critique of the analysis of the problems. In this manner, any disagreement about recommendations will lead to deeper understanding of the problems with current practice in the community, and to greater clarity about the community's values.

2. TOOLS AND TECHNOLOGIES FOR LANGUAGE DOCUMENTATION AND DESCRIPTION. Language documentation projects are relying more and more on new digital technologies and software tools. This section surveys a broad range of current practices, covering general-purpose software, specialized tools, and digital technologies. This snapshot of how digital language documentation and description are created and managed in practice provides a backdrop for our later analysis of data portability problems.

2.1. GENERAL PURPOSE TOOLS. Most computer-based language documentation work uses conventional office software. This software is readily available, often preinstalled, and familiar. Word processors have often been used in creating large dictionaries, such as a Yoruba lexicon with 30,000 entries split across twenty files (Yiwola Awoyale, p.c. 1998). Frequently cited benefits are WYSIWYG editing (i.e. 'what you see is what you get'), the find/replace function, the possibility of cut-and-paste to create sublexicons, and the ease of publishing. On the down side, a large fraction of the linguist's time is spent on maintaining consistency of both content and format or on finding ways to work around the lack of consistency. Word processors have also been used for interlinear text, with three main approaches: fixed-width fonts with hard spacing, manual setting of tab stops, and tables.[2] All methods require manual line-breaking, and

[2] http://www.linguistics.ucsb.edu/faculty/cu mming/WordForLinguists/Interlinear.htm

significant additional labor on presentation if line width or point size are ever changed. Another kind of office software, the spreadsheet, is often used for wordlists or paradigms.

Language documentation created using office software is normally stored in a proprietary format that is unsupported within five to ten years. While other export formats are supported, they may lose some of the structure. For instance, part of speech may be distinguished in a lexical entry through the use of italics, and this information may be lost when the data is exported to a nonproprietary plain-text format. Also, the portability of export formats may be compromised by being laden with presentational markup.

A second category of general-purpose software is hypertext processors. Perhaps the first well-known application to language documentation was the Macintosh hypercard stack that appeared in the late 1980s for *Sounds of the world's languages*, later published on the web[3] and on CD-ROM (Ladefoged 2000). More recently, the HTML standard coupled with universal, free browsers has encouraged the creation of large amounts of hypertext for a variety of documentation types. For instance, we have interlinear text with HTML tables (e.g. Peter Austin's Jiwarli fieldwork[4]), interlinear text with HTML frames (e.g. M. Eleanor Culley's presentation of Apache texts[5]), HTML markup for lexicons with hyperlinks from glossed examples and a thesaurus (e.g. Peter Austin and David Nathan's Gamilaraay lexicon[6]), gifs for representing IPA transcriptions (e.g. Steven Bird's Dschang tone paradigms[7]), and Javascript for image annotations (e.g. Bill Poser's annotated photographs of gravestones engraved with Déné syllabics[8]). In all these cases, HTML is used as the primary storage format, not simply as a view on an underlying database. The intertwining of content and format clearly makes this kind of language documentation difficult to maintain and reuse.[9]

The third category of general-purpose software is database packages. In the simplest case, the creator shares the database with others by requiring them to purchase the same package, and by shipping them a full dump of the database (e.g. the StressTyp database, which requires users to buy a copy of '4th Dimension'[10]). In other cases the dump is provided in a portable format, such as tab-delimited files or a set of SQL commands. A more popular approach is to host the database on a web-server and create a forms-based interface that allows remote users to search the database without installing any software (e.g. the Comparative Bantu Online Lexical Database[11] and the Maliseet-Passamaquoddy Dictionary[12]). Some databases support updates via the web (e.g. the Berkeley Interlinear Text Collector[13] and the Rosetta Project's site for uploading texts, wordlists, and descriptions[14]).

[3] http://hctv.humnet.ucla.edu/departments/li nguistics/VowelsandConsonants/

[4] http://www.linguistics.unimelb.edu.au/resear ch/projects/jiwarli/gloss.html

[5] http://etext.lib.virginia.edu/apache/ChiMe sc2.html

[6] http://coombs.anu.edu.au/WWWVLPages/Aborig Pages/LANG/GAMDICT/GAMDICT.HTM

[7] http://www.ldc.upenn.edu/sb/home/papers; shLDC2003S02

[8] http://www.ydli.org/dakinfo/dulktop.htm

[9] Our purpose in citing specific examples is not to single them out for criticism, but to show how serious work by conscientious scholars has grappled with a host of technical problems in the course of exploring a large space of imperfect solutions.

[10] http://www.let.leidenuniv.nl/ulcl/pil/stresstyp/

[11] http://www.cbold.ddl.ish-lyon.cnrs.fr/

[12] http://ultratext.hil.unb.ca/Texts/Malis eet/dictionary/index.html

[13] http://ingush.berkeley.edu:7012/BITC.html

[14] http://www.rosettaproject.org:8080/live/

2.2. SPECIALIZED TOOLS. Over the last two decades, several dozen tools with specialized support for language documentation and description have been developed; a representative sample is listed here.[15] Tools for linguistic data management include Shoebox[16] and the Fieldworks Data Notebook.[17] Speech analysis tools include Praat[18] and SpeechAnalyzer.[19] Many specialized signal annotation tools have been developed, including CLAN,[20] EMU,[21] and the Annotation Graph Toolkit[22] (including TableTrans, InterTrans and TreeTrans). There are many orthographic transcription tools, including Transcriber[23] and MultiTrans.[24] There are morphological analysis tools, such as the Xerox Finite State toolkit[25] and SIL's PC-Parse tools.[26] There are a wealth of concordance tools. Finally, some integrated multifunction systems have been created, such as LinguaLinks Linguistics Workshop.[27] The interested reader is referred to Antworth & Valentine 1998 for a full-length article on this topic.

In order to do their specialized linguistic processing, each of these tools depends on some model of linguistic information. All kinds of linguistic information—for example, time-aligned transcriptions, interlinear texts, syntax trees, lexicons—require suitable data structures and file formats. Given that most of these specialized tools have been developed in isolation, the models and formats are typically incompatible. For example, data created with an interlinear text tool cannot be subsequently annotated with syntactic information without losing the interlinear annotations. When interfaces and formats are open and documented, it is occasionally possible to cobble the tools together in support of a more complex need. However, the result is a series of increasingly baroque and decreasingly portable approximations to the desired solution. In consequence, specialized computational support for language documentation and description is in a state of disarray.

2.3. DIGITAL TECHNOLOGIES. A variety of digital technologies are used in language documentation owing to sharply declining hardware costs. These include technologies for digital signal capture (audio, video, physiological) and signal storage (hard disk, CD-R, DVD-R, DAT, minidisc).

Software technologies are also playing an influential role as new standards are agreed. At the micro level we have the simple hyperlink, which can connect linguistic descriptions to underlying documentation, for example, relating an analytical transcription to a recording. Hyperlinks streamline the descriptive process. Transcriptions can be checked with mouse clicks instead of unearthing an old tape or finding a speaker of the language. Hyperlinks help to organize the documentation, bringing temporally and

[15] Further examples may be found on SIL's page on *Linguistic computing resources* (http://www.sil.org/linguistics/computing.html) on the *Linguistic exploration* page (http://www.ldc.upenn.edu/exploration/), and on the *Linguistic annotation* page (http://www.ldc.upenn.edu/annotation/).

[16] http://www.sil.org/computing/shoebox/

[17] http://fieldworks.sil.org/

[18] http://fonsg3.hum.uva.nl/praat/

[19] http://www.sil.org/computing/speechtool s/speechanalyzer.htm

[20] http://childes.psy.cmu.edu/

[21] http://www.shlrc.mq.edu.au/emu/

[22] http://sf.net/projects/agtk/

[23] http://www.etca.fr/CTA/gip/Projets/ Transcriber/

[24] http://sf.net/projects/agtk/

[25] http://www.xrce.xerox.com/competencies/ content-analysis/fst/

[26] http://www.sil.org/computing/catalog/ pc-parse.html

[27] http://www.sil.org/LinguaLinks/LingWksh .html

spatially separated documentation together, and permitting a single artifact to play a role in multiple descriptions. This continual rearrangement of evidence is an important part of the analytic process.

At the macro level, software technologies and standards have given rise to the internet, which facilitates collaboration in the construction of language resources and low-cost dissemination of the results. Notably, it is portability problems that prevent the basic digital technologies from having their full impact. Thus, while the internet makes it easy to download a language resource, the would-be user may still face a daunting amount of set-up work before being able to derive the full benefits of that resource. The following download instructions for the Sumerian lexicon[28] illustrate the complexities (hyperlinks are underlined):

> Download the Sumerian Lexicon as a Word for Windows 6.0 file in a self-extracting WinZip archive.
> Download the same contents in a non-executable zip file.
> Includes version 2 of the Sumerian True Type font for displaying transliterated Sumerian. Add the font to your installed Windows fonts at Start, Settings, Control Panel, Fonts. To add the Sumerian font to your installed Windows fonts, you select File and Add New Font. Afterwards, make sure that when you scroll down in the Fonts listbox, it lists the Sumerian font. When you open the SUMERIAN.DOC file, ensure that at File, Templates, or at Tools, Templates and Add-Ins, there is a valid path to the enclosed SUMERIAN.DOT template file. If you do not have Microsoft's Word for Windows, you can download a free Word for Windows viewer at Microsoft's Web Site.
> Download Macintosh utility UnZip2.0.1 to uncompress IBM ZIP files. To download and save this file, you should have Netscape set in Options, General Preferences, Helpers to handle hqx files as Save to Disk. Decode this compressed file using Stuffit Expander.
> Download Macintosh utility TTconverter to convert the IBM format SUMERIAN.TTF TrueType font to a System 7 TrueType font. Decode this compressed file using Stuffit. Microsoft Word for the Macintosh can read a Word for Windows 6.0 document file. There is no free Word for Macintosh viewer, however.

The complexities illustrated in these download instructions are often encountered. Moreover, the ability of a technically savvy user to handle such complexities offers no guarantee that the software will actually work in that user's environment. For instance, the user could have a hardware or system software configuration that is substantially different than the one on which the resource was developed. Clearly our technologies for storing and delivering language resources fall far short of our need for easy reuse.

2.4. Digital archives. Recently several new digital archives of language documentation and description have been established, such as the Archive of the Indigenous Languages of Latin America,[29] and the Rosetta Project's Archive of 1000 Languages.[30] These exist alongside older archives that are in various stages of digitizing their holdings, for instance: the Archive of the Alaska Native Language Center,[31] the LACITO Linguistic Data Archive,[32] and the US National Anthropological Archives.[33] These archives and many others are surveyed on the *Language Archives* page.[34] Under the aegis of OLAC, the *Open Language Archives Community*,[35] the notion of language

[28] http://www.sumerian.org/

[29] http://www.ailla.org/

[30] http://www.rosettaproject.org/

[31] http://www.uaf.edu/anlc/

[32] http://lacito.vjf.cnrs.fr/archivage/

[33] http://www.nmnh.si.edu/naa/

[34] http://www.ldc.upenn.edu/exploration/ar chives.html

[35] http://www.language-archives.org/

archive has been broadened to include corpus publications by organizations like the Linguistic Data Consortium[36] and archives of linguistic software like the Natural Language Software Registry.[37]

Conventional language archives face many challenges, the most significant being the unfortunate reality that data preservation is not as attractive to sponsors as data creation. Other challenges may include: identifying, adapting, and deploying digital archiving standards; setting up key operational functions such as processing digital submissions, offsite backup, and migration to new digital formats and media over time; supporting new access modes (e.g. search facilities) and delivery formats (e.g. streaming media); and obtaining the long-term backing of an established institution that can credibly commit to providing preservation and access over the long term.

This survey, brief and incomplete as it is, makes clear that there is an abundance of tools and technologies for language documentation and description, and that the community is impressively adept at creating digital data. Yet the snapshot also reveals an embarrassing level of digital detritus. Expensive data cannot be reused, or else it requires a major recycling effort to salvage the valuable pieces.

Computers are not to blame for all problems of portability in language documentation and description, however; many portability problems predate the digital era. No earlier generation of linguists was able to be confident of discovering, accessing, and interpreting all relevant language resources. While the digital revolution has exacerbated some portability problems, particularly in such areas as format, citation, and preservation, it has simultaneously provided new, promising solutions to these older open problems, along with efficient processes for geographically dispersed communities to reach consensus about best practice. In the next section we consider an extensive set of portability problems under seven headings encompassing both digital and nondigital practices.

3. Seven problem areas for portability. During the rapid uptake of new digital technologies described in §2, many creators of language documentation and description have turned a blind eye to the issue of portability. Unfortunately, as a direct consequence of this, the fruits of their labors are likely to be unusable within five to ten years. In this section we identify seven problem areas for the portability of language documentation and description. While the tone of the discussion is negative, a full and frank assessment is necessary before we can articulate the core values that are being compromised by current digital and nondigital practices.

3.1. Content. By content we mean the information content of the resource. The area of content involves three key concepts: the breadth and depth of coverage, accountability for conclusions reached in description, and the terminology used in description.

coverage. The language documentation community has been active since the nineteenth century (even earlier in some cases[38]), collecting wordlists and texts, and writing descriptive grammars. With the arrival of the digital era, we can transfer the endeavor from paper to word processor and carry on as before. However, new technologies provide opportunities to create new kinds of language resources. We can make digital multimedia recordings of rich linguistic events, documenting endangered languages

[36] http://www.ldc.upenn.edu/

[37] http://registry.dfki.de/

[38] Celebrated early grammarians include Pāṇinī (5th century bc), Dionysius of Thrace (2nd century bc), and Hesychius of Alexandria (5th century ad).

and genres, and fortuitously capturing items that turn out to be crucial for later analysis. However, even when extensive multimedia recordings are made, they may be of low quality (e.g. poor microphone placement, bad lighting), or they may not represent a balanced collection (e.g. twenty recordings of the same genre). Each of these weaknesses in coverage limits our ability to interpret the content. Many senses, collocations, and constructions will be missed or else unique, and we will not have a corpus from which we can draw reliable conclusions.

ACCOUNTABILITY. The content of a description is difficult to verify when it cannot be checked against the language documentation on which it is based. For example, if the reported phonetic transcription of a word contradicts the known phonotactic properties of the language, could this be a typographical error, a difference in transcription practice, or a bona fide exception? Similarly, incompatible descriptions cannot be reconciled when the documentation is unavailable. Without accountability, problems of interpretation may only be resolved by contacting the author or by locating speakers of the same speech variety (and that only when the point in question does not derive from the idiosyncratic performance of the original source), and these problems may present significant obstacles to the reuse of the language description. Accountability is also an issue for documentation: heavy editing of recorded materials may give an artificial or even misleading impression of the original linguistic event.

TERMINOLOGY. Many potential users of language data are interested in assimilating multiple descriptions of a single language to gain an understanding of the language that is as comprehensive as possible. Many users are interested in comparing the descriptions of different languages in order to apply insights from one analysis to the analysis of another language, or to test a typological generalization. However, two descriptions may be difficult to compare or assimilate because they have used terminology differently.

Language documentation and description of all types depend critically on technical notation and vocabulary, and ambiguous or unknown terms compromise portability. For instance, the symbols used in phonetic transcription have variable interpretation depending on the descriptive tradition: ‘it is crucial to be aware of the background of the writer when interpreting an unexplained occurrence of [y]’ (Pullum & Ladusaw 1986:168). In morphosyntax, the term ‘absolutive’ can refer to one of the cases in an ergative language, or to the unpossessed form of a noun as in the Uto-Aztecan tradition (Lewis et al. 2001:151), and a correct interpretation of the term depends on an understanding of the linguistic context.

The existence of variable or unknown terms leads to problems for retrieval. Suppose that a linguistic typologist wanted to search the full-text content of a large collection of data from many languages in order to discover which languages have a particular trait. Since the terms are not standardized, the user will discover irrelevant documents (low precision) and will fail to discover relevant documents (low recall). In order to carry out a comprehensive search, the user must know all the ways in which a particular phenomenon is described. Even once a set of descriptions is retrieved, users will generally not be able to make reliable comparisons between the descriptions of different languages without studying them in detail. We will return to this topic when we discuss the problem of discovery.

3.2. FORMAT. By FORMAT we mean the manner in which the information is represented electronically. The area of format involves four key concepts: the OPENNESS of the format, the ENCODING of characters within textual information, the MARKUP of structure in the information, and the RENDERING of information in human-readable displays.

OPENNESS. Language data frequently ends up in a secret proprietary format. To use such data one must typically purchase commercial software from the company that developed the format, then install it on the same hardware and under the same operating system as used by the person who created the data. By contrast, an open format is one for which the specifications are open to the public, and thus software is available from multiple sources (including noncommercial ones) and on multiple platforms.

ENCODING. Encoding is the property of textual data that has to do with how the characters are represented as numerical codes in storage (as opposed to how they are keyboarded, or how they are rendered on the screen [Becker 1984]). This has been a perennial problem for linguists who need to encode characters that are not part of the standard character sets that are supported by common software, whether these be 'special' characters that occur in the orthographies of little-studied languages or symbols that are used in phonetic transcription. In the void left by the lack of standards, linguists have devised a variety of ingenious solutions, including using combinations of available characters to transliterate unavailable ones and devising new character sets that assign the needed characters to specific numerical codes. The portability of such solutions depends critically on the transmission of documentation that explains the encoding schemes. The emergence of Unicode[39] as a character encoding standard for all the major orthographic systems of the world (including the International Phonetic Alphabet) holds much promise. But even Unicode has portability problems when the characters that linguists need are not covered by the standard and they are forced to use the Private Use Area to encode custom characters.

MARKUP. Markup is the property of textual data that has to do with how the information above the character strings themselves is represented. For instance, in a dictionary the markup has to do with identifying the various parts of the dictionary entries. The purpose of markup is to support format conversion, database storage, and query. In a word processor, a linguist might switch fonts (such as from normal face to bold face) to indicate a particular part of the entry (such as the part of speech), as shown in 1a. This is the least portable markup of all, since such binary formatting can easily be lost when the file format is converted. Another approach to markup using a conventional word processor is for the linguist to use punctuation marks in a disciplined way (e.g. putting square brackets around the part of speech in a lexical entry), as shown in 1b. However, when maintaining complex entries it is easy to introduce a formatting error (e.g. omitting a closing bracket), with unpredictable consequences for the software used for converting, storing, or querying that data.

(1) a. *chien* **n** dog.
b. chien: [n] dog.

A more robust approach to markup is to introduce special strings of characters (called MARKERS or TAGS) into the stream of text. For instance, the Shoebox program uses markers that begin with a backslash to mark the beginnings of information elements, as shown in 2.

(2) \ent chien
\pos n
\def dog

An even more robust approach to markup uses balancing tags to mark both the beginning and end of each information element. Two examples are shown in 3. These follow

[39] http://www.unicode.org/

the markup convention first established in SGML, the Standard Generalized Markup Language,[40] of placing tags in angle brackets and using a slash within the tag to indicate the balancing end tag for the start tag of the same name.

(3) a. ⟨p⟩⟨font size = + 1⟩⟨i⟩chien⟨/i⟩⟨/font⟩

⟨b⟩n⟨/b⟩ ⟨font color = blue⟩dog.⟨/font⟩⟨/p⟩

b. ⟨entry⟩

⟨headword⟩chien⟨/headword⟩

⟨pos⟩n⟨/pos⟩

⟨definition⟩dog⟨/definition⟩

⟨/entry⟩

In markup systems there is a basic dichotomy between PRESENTATIONAL versus DESCRIPTIVE markup (Coombs et al. 1987). In presentational markup, the markup tags document what the information is supposed to look like (e.g. an entry is formatted as a paragraph with the headword in italics one font size larger, with the part of speech in bold, and with the definition in blue), as shown in 3a. This example uses HTML, Hypertext Markup Language, which is the most widely used system of presentational markup. It is portable with respect to preserving the appearance of information for human readers, but is not portable for the purpose of enabling computer systems to read the information and manipulate it consistently. For this, descriptive markup is needed in which the markup tags identify the pieces of information with respect to their function (e.g. an entry contains a headword, a part of speech, and a definition), as shown in 3b. This example illustrates XML,[41] Extensible Markup Language, which is now the most widely used system for implementing descriptive markup. The portability of descriptive markup may be limited when the system of markup is not documented. XML addresses this by supporting the formal definition of the markup scheme by means of a Document Type Definition (DTD) or an XML Schema (Bradley 2002).

RENDERING. It is a basic requirement of language resources that they should be presented to human readers in conventionally formatted displays (Simons 1998:§6). Both encoding and markup may lead to problems for rendering. Character encoding (the representation of characters in digital storage) causes problems for rendering when the fonts needed to view the textual information are not available. This problem is exacerbated when custom fonts are developed to support custom character sets. This is because the fonts themselves are a special kind of resource and are subject to a wide range of portability problems.

Markup may also cause problems for rendering. As we have seen, resources employ descriptive markup to maximize portability across computer systems and potential uses. However, such resources fail to cross the gap from computer to human if there is no meaningful way to display them.

3.3. DISCOVERY. By DISCOVERY we mean the problem of finding digital resources in the first place. The area of discovery involves two key concepts: discovering the EXISTENCE of a resource, and then judging the RELEVANCE of a discovered resource.

EXISTENCE. A given resource, even if it is of the highest quality, is of little practical value if the people who could benefit from it do not know that it exists. A large proportion of digital language resources (particularly those resulting from linguistic field work) are only to be found in the linguist's personal collection of computer files,

[40] http://xml.coverpages.org/sgml.html

[41] http://www.w3.org/XML/

and have no publicly available metadescription that would permit someone else to find them. When resources are entered into an institutional collection that is properly organized and cataloged, they remain virtually inaccessible if that catalog may only be consulted in person at the host institution. Even when a catalog is available electronically over the internet, the resources remain hidden unless the catalog is formatted in such a way that web search engines can appropriately retrieve its contents. In many of these cases, successful discovery of language resources depends on word-of-mouth and queries posted to electronic mailing lists.

RELEVANCE. Merely knowing that a resource exists is insufficient; the potential user must also be supplied with enough information in order to gauge the relevance of the resource. One may download a large resource only to discover that it is in an incompatible format. One may locate a binary file called dict.dat, then expend considerable effort to determine whether its content is relevant. Even where organized collections provide metadescription for subject language and linguistic type, they will typically use free text rather than a controlled vocabulary, reducing precision and recall in searching (cf. our discussion of terminology in §3.1).

3.4. ACCESS. By ACCESS we mean issues relating to the way in which the potential user of a resource gains access to it. Access involves three key concepts: the SCOPE of access that is granted, the PROCESS by which access is granted, and the EASE with which access is obtained.

SCOPE OF ACCESS. In the past, primary documentation was usually not disseminated. To listen to a field recording it was often necessary to visit the laboratory of the person who collected the materials, or to make special arrangements for the materials to be copied and posted. Digital publication on the web alleviates this problem, although projects usually refrain from full dissemination by limiting access to a restrictive search interface. This means that only selected portions of the documentation can be downloaded, and that all access must use categories predefined by the provider. Lack of full access means that materials are not fully portable.

PROCESS FOR ACCESS. It sometimes happens that an ostensibly available resource turns out not to be available after all, because there is no process whereby it may be obtained. One may discover the resource because its creator cited it in a manuscript or an annual research report. Commonly, researchers want to be recognized for the labor that went into creating primary language documentation, but do not want to make the materials available to others until they have derived maximum personal benefit. Despite its many guises, this problem has two distinguishing features: someone draws attention to a resource in order to derive credit for it—'parading their riches' as Mark Liberman (p.c., 2000) has aptly described it—and then applies undocumented or inconsistent restrictions to prevent access. The result may be frustration that a needed resource is withheld, leading to wasted effort or a frozen project, or to suspicion that the resource is defective and so must be protected by a smoke screen.

EASE OF ACCESS. Some resources are disseminated only on the web, making them difficult or impossible to access by people having a low-bandwidth connection or no connection at all. It may be particularly significant for communities that use endangered languages to have access to printed versions of language resources for use in efforts at language development and revitalization. In the case of multimedia resources, the absence of a low-bandwidth surrogate or a textual account of the content forces potential users to download and review the full resource in order to evaluate its suitability.

3.5. CITATION. By CITATION we mean the problems associated with making bibliographic citations of electronic language documentation and description. Citation involves four key concepts: the ability to cite a resource in a BIBLIOGRAPHY, the PERSISTENCE of electronic resource identifiers, the IMMUTABILITY of materials that are cited, and the GRANULARITY of what may be cited.

BIBLIOGRAPHY. Research publications are normally required to provide full bibliographic citation of the materials used in conducting the research. Citation standards are usually high when citing conventional publications, but are much lower for citations of digital language resources. Many scholars do not know how to cite electronic resources; thus the latter are often incorrectly cited, or not cited at all.[42] When electronic sources are not properly cited, it is difficult to discover what resources were used in conducting the research or, following the linkage in the reverse direction, to consult a citation index to discover all the ways in which a given resource has been used.

PERSISTENCE. Often a language resource is available on the web, and it is convenient to identify the resource by means of its UNIFORM RESOURCE LOCATOR (URL) since this may offer the most convenient way to obtain the resource. However, URLs are notorious for their lack of persistence. They 'break' when the resource is moved or when some piece of the supporting infrastructure, such as a database server, ceases to work.

IMMUTABILITY. Even if a URL does not break, the item that it references may be mutable, changing over time. Language resources published on the web are usually not versioned, and a third-party description based on some resource may cease to be valid if that resource is changed. This problem can be solved by archiving each version and ensuring that citations reference a particular version. Publishing a digital artifact, such as a CD, with a unique identifier, such as an ISBN, also avoids this problem.

GRANULARITY. Citation goes beyond bibliographic citation of a complete item. We may want to cite some component of a resource, such as a specific narrative or lexical entry. However, the format of the resource may not support durable citations to internal components. For instance, if a lexical entry is cited by a URL that incorporates its lemma, and if the spelling of the lemma is altered, then the URL will not track the change. In sum, the portability of a language resource suffers when incoming and outgoing links to related materials are fragile.

3.6. PRESERVATION. By PRESERVATION we mean the problem of ensuring that digital resources remain accessible to future generations. Preservation involves three key concepts: the LONGEVITY of the format, the SAFETY of resources from catastrophic loss, and the ongoing migration of resources to current physical and digital MEDIA.

LONGEVITY. The digital technologies used in language documentation and description greatly enhance our ability to create data while simultaneously compromising our ability to preserve it. Compared to paper copy, which can survive for hundreds of years, and other media such as clay tablets, which have lasted for millenia, digitized materials are evanescent because they are based on binary formats. The problem is exacerbated when they use a proprietary format that becomes obsolete within a few years (e.g. Microsoft Word 3.0). Presentational markup with HTML and interactive content with CGI, Javascript, and specialized browser plugins require future browsers and servers to be backwards-compatible. Worse still, primary documentation may be embodied in the interactive behavior of the resource (e.g. the gloss of the text under the mouse may show up in the browser status line, using the Javascript 'mouseover' effect). Consequently, digital resources—especially dynamic or interactive ones—often have a short lifespan, and typically become unusable three to five years after they cease to be actively maintained.

[42] Incidentally, *The Columbia guide to online style* (Walker & Taylor 1998) is a good source on how to cite online resources.

SAFETY. Language resources are stored on some physical medium or device (the CARRIER), such as paper, magnetic tape, and various kinds of disk (e.g. floppy disk, hard drive, compact disk). Many undesirable eventualities may befall such physical artifacts; they may be degraded, damaged, lost, stolen, or destroyed. Such problems are usually greater in the field, where accidents may be more common (e.g. canoes capsizing), and where there may be less protection from extremes of climate. If the resource is digital it may be deleted, overwritten, or corrupted. While the individual guardian of the resource may exercise great care with it, mistakes nevertheless occur. Other agents also come into play: the people who share, manage, or repair the equipment; hostile third parties including thieves and computer viruses; political instability that may force sudden evacuation; elements of the environment such as dust, humidity, pests, mold, and power failure; catastrophes including fire, flood, lightning strike, and war; and natural disasters such as earthquakes, tornadoes, hurricanes, tsunamis, and volcanic eruptions. A resource may suddenly cease to exist if no steps are taken to mitigate these risks by ensuring that another copy is in a safe location.

MEDIA. Digital storage media may become inaccessible due to the absence of supporting hardware (e.g. 5.25'' floppy disks). While the problem of obsolete media predates the digital era (e.g. wax cylinder recordings), the problem has become more acute and is frequently noted in recent literature on digital archives: 'The lifespan of consumer physical digital media is estimated to be 5 years or less' (Cohen 2001); 'To date, none of the digital recording systems developed specifically for audio has achieved a proven stability in the market place, let alone in an archive' (International Association of Sound and Audiovisual Archives 2001). Magnetic media degrade in quality over time, with the loss of signal strength and, in the case of tapes, deformation of the backing, hydrolysis in the binder (St-Laurent 1996), and the imposition of bleedthrough.

3.7. RIGHTS. By RIGHTS we mean issues relating to what a potential user of a resource is permitted to do with the resource. The area of rights involves four key concepts: clarifying the TERMS OF USE for the resource, maximizing the public BENEFIT of the resource, protecting any SENSITIVITY that is inherent in the resource, and finding the proper BALANCE between public benefits and protecting sensitivities.

TERMS OF USE. A variety of individuals and institutions may have intellectual property vested in a language resource, and there is a complex terrain of legal, ethical, and policy issues involved (Liberman 2000). In spite of this, most digital language data is disseminated without identifying the copyright holder and without any license delimiting the range of acceptable uses of the material. Often people collect or redistribute materials or create derived works without securing the necessary permissions. While this is often benign (e.g. when the resources are used for research purposes only), the creator or user of the resource risks legal action, or having to restrict publication, or even having to destroy primary materials. To avoid any risk one must avoid using materials whose property rights are in doubt. In this way, the very lack of documented rights may restrict the portability of the language resource.

Sometimes resources are not made available on the web for fear that they will get into the wrong hands or be misused. However, this fear may be based on a confusion between dissemination medium and rights. The web supports secure data exchange between authenticated parties through data encryption. Copyright statements and user licenses can restrict uses. More sophisticated models for managing digital rights are emerging (Iannella 2001). The application of these techniques to language resources is unexplored, and today we have an all-or-nothing situation in which the existence of any restriction tends to prevent access across the board.

BENEFIT. Researchers typically want the results of their work to benefit human knowledge and experience as widely as possible. When permission is obtained for collecting primary language documentation, however, restrictions may be imposed on who is allowed to use it, how they are allowed to use it, and the time period of use. Such restrictions may originate from various sources including the language community, the government agency that provides the research permit, or an institutional review board. While researchers may wish the results of their work to benefit the public, they may discover too late that legitimate but unanticipated uses by unforeseen users are unintentionally jeopardized when permissions are tightly circumscribed.

SENSITIVITY. Many individuals and institutions are sensitive about the collection, dissemination, and uses of what linguists typically regard as neutral language documentation. The content of an oral discourse may contain sensitive personal, tribal, religious, or corporate information, or may be viewed as libel, breach of confidence, or even treason by others. There is a perceived risk of commercial exploitation of language documentation that, as with Western commercialization of indigenous music, may 'emphasize the exotic and the unexpected at the expense of the real substance' (Bebey 1997:1). Researchers may build a career on the data obtained from a language community without ever making the resources available in a form that benefits that community. Disregard for such sensitivities may compromise the standing or security of an individual or group, or may lead to the imposition of tighter access restrictions in the future (Wilkins 1992).

BALANCE. Access restrictions that protect a sensitive resource simultaneously limit the wider benefit that the resource may bring to human knowledge and experience. Researchers will typically want to maximize the wider benefit of the resource while protecting any sensitivities. The precise formulation of access restrictions, however, is often overgeneral, encompassing a greater timespan or a greater proportion of the resource than strictly necessary. It causes real problems when a sensitivity is stipulated without any time limit. An item that could never be accessed (including at no time in the future) would only be wasting space in an archive. The sensitivities inherent in a resource are often time-limited, for example, by the lifetime of the individuals involved in creating it, or the remaining lifetime of an endangered language. Sometimes, sensitivities that pertain to some part of the linguistic documentation are assigned scope over an entire collection. For instance, when a portion of a video recording contains some sensitive material this may constitute grounds for withholding the entire recording. The sensitivity may be generalized from a recording to the associated linguistic description, such as transcripts, even though the transcripts themselves may contain no sensitive material. In the reverse direction it is also possible for sensitivities about the linguistic description to be generalized to the underlying documentation. The researcher may not be prepared to release the primary documentation until satisfied with the transcriptions, on the grounds that his or her career will benefit more if he or she has sole access to the primary documentation while conducting the research.

3.8. SPECIAL CHALLENGES FOR LITTLE-STUDIED LANGUAGES. Many of these problems are exacerbated in the case of little-studied languages. The small amount of existing work on the language and the concomitant lack of established documentary practices and conventions may lead to especially diverse nomenclature. Inconsistencies within or between language descriptions may be harder to resolve because of the lack of significant documentation, the limited access to speakers of the language, and the limited understanding of dialect variation. Open questions in one area of description (e.g.

the inventory of vowel phonemes) may multiply the indeterminacies in another (e.g. the transcription and interpretation of texts). More fundamentally, existing documentation and description may be virtually impossible to discover and access, owing to its sparse or fragmentary nature.

The acuteness of these portability problems for little-studied languages can be highlighted by comparison with well-studied languages. In English, published dictionaries and grammars exist to suit all conceivable tastes, and it therefore matters little, relatively speaking, if some of these resources are not especially portable. However, when there is only one available dictionary for a little-studied language, it must be pressed into a great range of services, and so portability becomes a major concern.

Another issue that is more vexing in the case of endangered languages is access. Access may be prevented by the choice of inappropriate media for dissemination. For instance, an endangered language dictionary published only on the web will not be accessible to speakers of that language who live in a village without electricity. In the reverse direction, when a collection of recordings is transcribed in a little-studied language but not interpreted into a major language, then the content of those recordings is inaccessible to the outside world.

Sensitivity issues are often more acute for endangered languages. The wishes of the speech community (to control rather than disseminate their language) may conflict with the wishes of the linguists documenting the language (to disseminate rather than tie up the documentation). In balancing sensitivities it is often helpful to distinguish description from documentation; researchers *create* descriptions, while they only *collect* documentation. In the case of pure documentation, such as a video recording of a linguistic event in which the researcher has no creative input, the sensitivity of the participants takes precedence over any sensitivities of the researcher. In the case of pure description, such as a theoretical monograph on the language, the researcher's own sensitivities prevail. However, language resources such as grammars and analytical lexicons combine documentation and description. In such cases, resolving the conflicting sensitivities of the speech community and the linguists documenting the language will often depend on forging alliances and establishing shared goals.

This concludes our discussion of the portability problems in language documentation and description. The following sections respond to these problems by laying out the core values that constitute requirements for best practices (§4), describing how the the Open Language Archives Community supports the process of identifying community-agreed best practices (§5), and by providing a comprehensive set of best practice recommendations (§6).

4. Value statements. Best practice recommendations amount to a decision about which of several possible practices is best. As anthropologist Henry Bagish points out in his critique of cultural relativism, indiscriminate tolerance of every possible practice is paralyzing (Bagish 1983). He proposes a formula that permits objective, crosscultural evaluation of competing practices, namely, 'If you value X, then A is better than B'. That is, before making a judgment as to which practice is better, one must clearly articulate the values that motivate the choice. If different parties can agree on the motivating values, then they should be able to come to agreement on the evaluation of competing practices.

In this section, we articulate the values that motivate the recommendations for best practice that are offered in §6. Our use of 'we' in the value statements is meant to include readers and members of the wider language resources community who share these values. Note that these statements do not necessarily reflect an official position of the Linguistic Society of America.

4.1. Content. coverage. We value comprehensive documentation, especially for little-studied languages. Thus the best practice is one that establishes a record that is sufficiently broad in scope, rich in detail, and authentic in portrayal that future generations will be able to experience and study the language, even if no speakers remain.

accountability. We value the ability of researchers to verify language descriptions. Thus the best practice is one that provides the documentation that lies behind the description.

terminology. We value the ability of users to compare two resources by virtue of their terminology. Thus the best practice is one that makes it easy to identify the comparable aspects of unrelated resources.

4.2. Format. openness. We value the ability of any potential user to make use of a language resource without needing to obtain unique or proprietary software. Thus the best practice is one that puts data into a format that is not proprietary.

encoding. We value the ability of users of a resource to understand the textual characters that are used in the resource, even in the absence of a font that can correctly render them. Thus the best practice is one that fully documents what the character codes in the resource represent.

markup. We value the ability of users of a resource to be able to write programs that can process or present the information in novel ways. Thus the best practice is one that represents all of the information using a transparent descriptive markup, rather than in procedural code or in presentational markup.

rendering. We value the ability of users of a resource to be able to read the content of the information in a conventional presentation form. Thus the best practice is one that supplements the information resource with all the auxiliary software resources that are needed to render it for display.

4.3. Discovery. existence. We value the ability of any potential user of a language resource to learn of its existence. Thus the best practice is one that makes it easy for anyone to discover that a resource exists.

relevance. We value the ability of potential users of a language resource to judge its relevance without first having to obtain a copy. Thus the best practice is one that makes it easy for anyone to judge the relevance of a resource based on its description.

4.4. Access. scope of access. We value the ability of any potential user of a language resource to access the complete resource, not just a limited portion of it or a limited interface to it. Thus the best practice is one that makes it easy for users to obtain a complete copy of the resource.

process for access. We value the ability of any potential user of a language resource to follow a well-defined procedure to obtain a copy of the resource. Thus the best practice is one in which there is a clearly documented procedure by which users may obtain a copy of the resource.

ease of access. We value the ability of potential users to access a version of a language resource from wherever they are located, even where the available computational infrastructure may be limited. Thus the best practice is one that makes such access possible.

4.5. Citation. bibliography. We value the ability of users of a resource to give credit to its creators, as well as to learn the provenance of the sources on which it is based. Thus the best practice is one that makes it easy for electronic language documentation and description to be cited.

PERSISTENCE. We value the ability of users of language resources to locate an instance of the resource, even though its actual location or filename might change. Thus the best practice is one that archives resources with identifiers that are independent of location or file name.

IMMUTABILITY. We value the ability of users to cite a language resource without that resource changing and invalidating the citation. Thus the best practice is one that makes it possible for users to cite particular versions that never change.

GRANULARITY. We value the ability of potential users to cite the component parts of a language resource. Thus the best practice is one that ensures each subitem of a resource has a durable identifier.

4.6. PRESERVATION. LONGEVITY. We value ongoing access to language resources over the very long term. Thus the best practice is one that stores resources in formats that are likely to remain usable for generations to come.

SAFETY. We value ongoing access to language resources over the very long term. Thus the best practice is one that stores copies of resources in multiple locations so as to ensure against catastrophic damage to a single repository.

MEDIA. We value ongoing access to language resources beyond the life span of any particular storage medium. Thus the best practice is one that migrates resources to new physical and digital media before the ones they are stored in become unusable.

4.7. RIGHTS. TERMS OF USE. We value the ability of potential users of a language resource to understand any restrictions on its permissible use before they begin to use it. Thus the best practice is one that clearly states the terms of use as part of the resource package.

BENEFIT. We value the maximal application of language resources toward the benefit of human knowledge and experience. Thus the best practice is one that does not hinder the fair use of a language resource for scientific, educational, humanitarian, or other noncommercial uses.

SENSITIVITY. We value the rights of the contributors to a language resource. Thus the best practice is one that protects any sensitivities stipulated by the contributors.

BALANCE. We value the potential long-term benefits of a resource, even when sensitivities prevent its dissemination in the near term. Thus the best practice is one that clearly identifies the nature of a sensitivity and associates it with an explicit time frame.

These value statements lead us to propose the detailed best-practice recommendations listed in §6. Before proceeding to these recommendations we give a brief overview of OLAC, which provides structures to support the elaboration and implementation of such recommendations.

5. OLAC, THE OPEN LANGUAGE ARCHIVES COMMUNITY. While this article sketches a set of values and practices designed to enhance the portability of digital language documentation and description, it is ultimately the community that must work out the details and reach a consensus. A community that can fill this role has already begun to form.

In December 2000, an NSF-funded workshop, Web-Based Language Documentation and Description, was held in Philadelphia. The workshop brought together a group of nearly 100 language software developers, linguists, and archivists who are responsible for creating language resources in North America, South America, Europe, Africa, the

Middle East, Asia, and Australia (Bird & Simons 2000). The outcome of the workshop was the founding of the Open Language Archives Community (OLAC),[43] with the following purpose:

> OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources.

Today OLAC has over twenty participating archives in seven countries, with over 30,000 records describing language resource holdings. The OLAC gateway at the LINGUIST List site[44] permits users to search the contents of all archives from a single location, before being directed to the website of the individual archive for information about how to obtain the resource. Anyone in the wider linguistics community can participate, not only by using the search facilities, but also by documenting their own resources, or by helping create and evaluate new best practice recommendations.

The technical infrastructure for OLAC is built on a framework developed within the digital libraries community by the Open Archives Initiative.[45] It has two components: a metadata standard (DCMI 1999) and a metadata harvesting protocol (Lagoze et al. 2002). These standards define how data providers—the institutions that want to make their resources known—publish metadata about their holdings, and how service providers—the institutions that want to provide value-added services for an entire community—can harvest the metadata and add it to the information pool on which they base their services. The OLAC versions of these standards, namely the OLAC Metadata standard and the OLAC Repositories standard, are designed to address the particular needs of language archiving (Bird & Simons 2003, Simons & Bird 2003a,b).

'Metadata' is structured data about data—descriptive information about a physical object or a digital resource. Library card catalogs represent a well-established type of metadata, and they have served as collection management and resource-discovery tools for decades. The OLAC Metadata standard (Simons & Bird 2002a) defines the elements to be used in metadata descriptions of language archive holdings, and how such descriptions are to be disseminated using XML descriptive markup for harvesting by service providers in the language-resources community. The OLAC metadata set contains the fifteen elements of the Dublin Core metadata set (DCMI 1999), plus several refined elements that capture information of special interest to the language-resources community. In order to improve recall and precision when searching for resources, the standard also defines a number of controlled vocabularies for descriptor terms. The most important of these is a standard for identifying languages (Simons 2000).

The OLAC Repositories standard (Simons & Bird 2002c) defines the protocol by which service providers query web-accessible repositories to harvest the metadata records they publish. Any other site may use the protocol to collect metadata records in order to provide a service, such as offering a union catalog of all archives or a specialized search service pertaining to a particular topic. To facilitate widespread discovery of the resources held in OLAC archives, all OLAC metadata is mapped to the more general-purpose Dublin Core metadata set and disseminated to the broader community of digital libraries; it is also mapped to an HTML format to facilitate indexing by web search engines. In the same way, more specialized metadata formats, such as the IMDI

[43] http://www.language-archives.org/

[44] http://www.linguistlist.org/olac/

[45] http://www.openarchives.org/

format for fine-grained description of linguistic field recordings,[46] can be mapped to OLAC metadata for dissemination to the wider language-resources community.

In addition to this technical infrastructure, OLAC also provides simple infrastructure to support interaction among the human participants of the Open Language Archives Community. The OLAC Process standard (Simons & Bird 2002b) defines: (i) the governing ideas of OLAC, including a summary statement of its purpose, vision, and core values; (ii) the organization of OLAC, in terms of the groups of participants that play key roles: coordinators, advisory board, council, participating archives and services, working groups, and participating individuals; and (iii) the operation of OLAC, in terms of a document process that defines how documents are generated and how they progress from one status to the next along the five-phase life cycle of development, proposal, testing, adoption, and retirement.

This last aspect of the OLAC Process (i.e. the document process) is already leading to new standards and best practice recommendations. In the future, we envision best practices for a variety of players, including linguists, archivists, developers, and sponsors. By participating in the OLAC Process—setting up working groups, reviewing current practices, formulating best practice recommendations, and forging a consensus in the wider community through cycles of review and revision—the community that creates and uses digital language documentation and description will move forward to a new era of highly portable language resources.

Having described suitable community infrastructure for developing best practice recommendations, we now present our own recommendations. By presenting them here we do not intend to bypass the consensus-building process, but rather to stimulate widespread discussion leading to better, more carefully articulated recommendations.

6. Best practice recommendations.

This section recommends best practices in support of the values set out in §4. These guidelines need to be fleshed out in more detail by the language-resources community. Note that these statements do not necessarily reflect an official position of the Linguistic Society of America.

6.1. Content.

(1) coverage.
- a. Make rich records of rich interactions, especially in the case of endangered languages or genres.
- b. Document the 'multimedia linguistic field methods' that were used.

(2) accountability.
- a. Provide the full documentation on which language descriptions are based. For instance, a grammar is based on a text corpus.
- b. When texts are transcribed, provide the primary recording (without segmenting it into clips).
- c. Transcriptions should be time-aligned to the underlying recording in order to facilitate verification.
- d. When recordings have been significantly edited, provide the original recordings to guarantee authenticity of the materials.

(3) terminology.
- a. Map the terminology and abbreviations used in description to a common ontology of linguistic terms.

[46] http://www.mpi.nl/world/ISLE/documents/draft/ISLE_MetaData_2.5.pdf

b. Map the element tags used in descriptive markup to a common ontology of linguistic terms.

c. Map the symbols used in transcription to phonological descriptors that are mapped to a common ontology of linguistic terms.

6.2. Format.

(4) openness.

a. Store all language documentation and description in formats that are open (i.e. whose specifications are published and nonproprietary).

b. Prefer formats supported by software tools available from multiple suppliers.

c. Prefer formats with free tools over those with commercial tools only.

d. Prefer published proprietary formats, e.g. Adobe Portable Document Format (PDF) and MPEG-1 Audio Layer 3 (MP3), to secret proprietary formats, e.g. Microsoft formats.

(5) encoding.

a. Encode the characters with Unicode.

b. Avoid Private Use Area characters, but if they are used, document them fully.

c. Document any 8-bit character encodings.

d. Document any scheme used to transliterate characters.

(6) markup.

a. Prefer descriptive markup over presentational markup.

b. Prefer XML (with an accompanying DTD or Schema) over other schemes of descriptive markup.

c. If the XML DTD or Schema is not a previously archived standard, archive it. Give each version a unique identifier.

d. If a descriptive markup scheme other than XML is used, prepare and archive a document that explains the markup scheme.

e. When a resource using descriptive markup is archived, reference the resource to the archived version of the definition of the associated markup format.

f. If punctuation and formatting are used to represent the structure of information, document how they are used.

(7) rendering.

a. If the fonts needed to appropriately render the resource are not commonly available, archive them and reference the resource to the archived version of the needed fonts.

b. Provide one or more human-readable versions of the material, using presentational markup (e.g. HTML) or other convenient formats. Proprietary formats are acceptable for delivery as long as the primary documentation is stored in a nonproprietary format.

c. If you have used stylesheets to render the resource, archive them as well.

N.B. Format is a critical area for the definition of best practices. We propose that recommendations in this area be organized by type (e.g. audio, image, text), possibly following the inventory of types identified in the Dublin Core metadata set.[47]

[47] http://dublincore.org/documents/dcmi-type-vocabulary/

6.3. Discovery.

(8) existence.

a. List all language resources with an OLAC repository.

b. Any resource presented in HTML on the web should contain metadata with keywords and description for use by conventional search engines.

(9) relevance.

a. Follow the OLAC recommendations on best practice for describing language resources using metadata, especially concerning language identification and linguistic data type. This will ensure the highest possibility of discovery by interested users in the OLAC union catalog hosted on the LINGUIST List site.[48]

6.4. Access.

(10) scope of access.

a. Publish complete primary documentation, providing a documented method by which anyone may obtain the documentation.

b. Publish documentation and description in such a way that users can gain access to the files to manipulate them in novel ways. (That is, do not just publish through a fixed user interface like a web search form, or a fixed presentation view like a PDF file.)

c. Transcribe all recordings in the orthography of the language (if one exists).

(11) process for access.

a. Document the process for access as part of the metadata, including any licenses and charges.

b. Document all restrictions on access as part of the metadata.

c. For resources not distributed over the web, document the expected delivery time.

d. For resources not distributed over the web, publish online surrogates that are easy for potential users to access and evaluate.

(12) ease of access.

a. Publish digital resources using appropriate delivery media, e.g. web for small resources, and CD or DVD for large resources.

b. Provide low-bandwidth surrogates for multimedia resources, e.g. publish MP3 files corresponding to large, uncompressed audio data.

c. Provide transcriptions for extended recordings to facilitate access to the relevant section.

d. For little-studied languages where the speech community has limited web access, publish print versions to facilitate access by the community, and provide a written account of any multimedia content using a major language.

6.5. Citation.

(13) bibliography.

a. Furnish complete bibliographic data in the metadata for all language resources created.

b. Provide complete citations for all language resources used.

c. Provide instructions on how to cite an electronic resource from the collec-

[48] http://www.linguistlist.org/olac/

tion as part of the web site for a digital archive (e.g. see the instructions on SIL's Electronic Working Papers site[49]).

d. Use the metadata record of a language resource to document its relationship to other resources (e.g. in the OLAC context, use the RELATION element).

(14) PERSISTENCE.

a. Ensure that resources have a persistent identifier, such as an ISBN, an OAI identifier, or a Digital Object Identifier.[50]

b. Ensure that a persistent identifier resolves to an online instance of the resource, or else to detailed online information about how to obtain the resource.

(15) IMMUTABILITY.

a. Provide fixed versions of a resource, either by publishing it on a read-only medium, or by submitting it to an archive that ensures immutability.

b. Distinguish multiple versions with a version number or date, and assign a distinct identifier to each version.

(16) GRANULARITY.

a. Provide a formal means by which the components of a resource may be uniquely identified.

b. Take special care to avoid the possibility of ambiguity, such as arises when lemmas are used to identify lexical entries, and where multiple entries can have the same lemma.

6.6. PRESERVATION. Many organizations have published detailed recommendations concerning the archival preservation of paper, audio, video, and images. Readers are referred to: the Library of Congress Preservation Directorate[51] which has recommendations concerning paper and images (Library of Congress 1995, 2001); the UNESCO Archives Portal[52] which has a section on preservation and conservation, including a reader on audiovisual archives focusing on the practical needs of audiovisual archivists in developing countries (Harrison 1997); the International Association of Sound and Audiovisual Archives[53] which has published recommendations for audio preservation (International Association of Sound and Audiovisual Archives 2001); The Council on Library and Information Resources[54] which publishes a series of reports containing chapters on audio and video preservation (Brylawski 2002, Cohen 2001, Wactlar & Christel 2002); The Conservation Online (CoOL) website,[55] with the most comprehensive set of links to online resources for the preservation of audio materials,[56] and recommendations for the handling of media (St-Laurent 1996); the Preservation Metadata Working Group of the Online Computer Library Center,[57] the Research Libraries

[49] http://www.sil.org/silewp/citation.htm l

[50] http://www.doi.org/

[51] http://lcweb.loc.gov/preserv/

[52] http://www.unesco.org/webworld/portal_archives/pages/

[53] http://www.iasa-web.org/

[54] http://www.clir.org/

[55] http://palimpsest.stanford.edu/

[56] http://palimpsest.stanford.edu/bytopic/ audio/

[57] http://www.oclc.org/research/pmwg/

Group,[58] developing a standard to 'document and evaluate the processes that support the long-term retention and accessibility of digital content' (OCLC/RLG 2002); and the International Standards Organization, providing a standard concerning the structure and function of a digital archive in ISO 14721 *Reference Model for an Open Archival Information System*.[59]

The recommendations in this section touch on key themes from the literature we cite that are directly relevant to language archiving. However, readers are advised to consult the literature for full discussion and detailed recommendations.

(17) LONGEVITY.

a. Commit all documentation and description to a digital archive that can credibly promise long-term preservation and access.

b. Ensure that the archive satisfies the key requirements of a well-founded digital archive, for instance, that it implements digital archiving standards, provides offsite backup, migrates materials to new formats and media/ devices over time, is committed to supporting new access modes and delivery formats, has long-term institutional support, and has an agreement with a national archive to take materials if the archive folds.

c. Digitize analog recordings, to permit lossless copying in the future.

d. Publish language documentation and description on the web using standard open formats so that they are fortuitously captured by internet archives (e.g. the Wayback Machine[60]).

e. When digital language resources are stored offline, transfer them to new storage media before the existing media type becomes unsupported (for many media types this would be necessary every five years).

f. Archive physical versions of the language documentation and description (e.g. printed versions of documents, any tapes from which online materials were created).

g. Prefer the file formats—including markup and encoding—that have the best prospect for accessibility far into the future (e.g. use type 1 (scalable) fonts in preference to bitmap fonts in documents).

(18) SAFETY.

a. Ensure that copies of archived documentation and description are kept at multiple locations (e.g. following the LOCKSS concept, 'Lots of copies keeps stuff safe'[61]).

b. Create a disaster recovery plan, such as that developed by the Syracuse University Library (1995), containing procedures for salvaging archived resources in the event of a disaster.

(19) MEDIA.

a. Whenever possible, maintain language resources on digital mass-storage systems, for easy backup and transfer to upgraded hardware.

b. Refresh offline digital storage by transferring the data to new storage at regular intervals (e.g. 1–5 years). Choose intervals appropriate for the performance of the media and location (e.g. offline magnetic media suffer

[58] http://www.rlg.org/

[59] http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

[60] http://www.archive.org/

[61] http://lockss.stanford.edu/

from signal loss and bleedthrough and should normally be refreshed every 1–2 years; nonmagnetic media and media maintained in climate-controlled storage may only need refreshing after 5–10 years).

c. Language resources that are stored in a proprietary binary format should be migrated to new formats before the existing format becomes unsupported (for many formats this would be necessary every five years).

6.7. RIGHTS.

(20) TERMS OF USE.

a. Ensure that the intellectual property rights relating to the resource are fully documented.

b. Ensure that there is a terms-of-use statement that clearly states what a user may and may not do with the materials.

(21) BENEFIT.

a. Ensure that the resource may be used for research purposes.

b. Ensure that the use of primary documentation is not limited to the researcher, project, or agency responsible for collecting it.

(22) SENSITIVITY.

a. Ensure that the nature of any sensitivity is documented in detail. To aid interpretation in the distant future, include concrete examples of any eventualities that must be avoided.

(23) BALANCE.

a. Limit any stipulations of sensitivity to the sensitive sections of the resource, permitting nonsensitive sections to be disseminated more freely.

b. Associate each sensitivity with an expiry date or a review date. List objective criteria that can be applied to determine whether the sensitivity has expired.

c. When primary documentation is closed in order for a researcher to derive maximal personal benefit, the expiry date should be no later than five years after the recording date.

As stated at the outset, we have structured this article to build consensus. Readers who take issue with any of our best-practice recommendations are encouraged to join the OLAC community[62] and enter into the consensus-building process. We further recommend that they review the corresponding statements of problems (§3) and values (§4). Baseline consensus on the problems and values provides a secure foundation for constructive discussions about the community's best practices.

7. CONCLUSION. Today, the community of scholars engaged in language documentation and description is in the midst of transition between the paper-based era and the digital era. We are still working out how to preserve knowledge that is stored in digital form. During this transition period, we observe unparalleled confusion in the management of digital language documentation and description. A substantial fraction of the resources being created can only be reused on the same software/hardware platform, within the same scholarly community, for the same purpose, and then only for a period of a few years. However, by adopting a range of best practices, this specter of chaos can be replaced with the promise of easy access to highly portable resources.

Using TOOLS as our starting point, we described a diverse range of practices and discussed their negative implications for DATA portability along seven dimensions, lead-

[62] http://www.language-archives.org/

ing to a collection of ADVICE on creating portable resources. These three categories, tools, data, and advice, are three pillars of the infrastructure provided by OLAC, the Open Language Archives Community (Bird & Simons 2001). Our best-practice recommendations are preliminary, and we hope they will be fleshed out by the community using the OLAC Process.

We leave off where we began, namely with tools. It is the community's use of new tools that has led to data portability problems. And it is only newer tools—supporting the kinds of practices we advocate—that will address these problems. An archival format is useless unless there are tools for creating, managing, and browsing the content stored in that format. Needless to say, no single organization has the resources to create the necessary tools, and no third-party developer of general-purpose office software will address the specialized needs of the language documentation and description community. We need nothing short of an open source[63] revolution, leading to new open source tools based on agreed data models for all of the basic linguistic types, connected to portable data formats, with all data housed in a network of interoperating digital archives. On their own, technological solutions will be inadequate, as they have been in the past, only contributing further to the digital carnage we experience today. Instead, the technological solutions must be coupled with a sociological innovation, one that produces broad consensus about the design and operation of common digital infrastructure for the archiving of language documentation and description.

REFERENCES

ANTWORTH, EVAN, and J. RANDOLPH VALENTINE. 1998. Software for doing field linguistics. In Lawler & Aristar Dry, 170–96. Appendix online: http://www.sil.org/computing/routledge/antworth-valentine/.

BAGISH, HENRY H. 1983. Confessions of a former cultural relativist. Anthropology annual editions 83/84, ed. by Elvio Angeloni, 22–29. Guilford, CT: Dushkin Publishing Group.

BEBEY, FRANCIS. 1997. African music: A people's art. Trans. by Josephine Bennett. New York: Lawrence Hill and Company.

BECKER, JOSEPH D. 1984. Multilingual word processing. Scientific American 251.96–107.

BIRD, STEVEN, and GARY SIMONS (eds.) 2000. Proceedings of the workshop on web-based language documentation and description. Online: http://www.ldc.upenn.edu/exploration/expl2000/.

BIRD, STEVEN, and GARY SIMONS. 2001. The OLAC metadata set and controlled vocabularies. Proceedings of the ACL/EACL workshop on sharing tools and resources for research and education, compiled by Mike Rosner, 27–38. East Stroudsburg, PA: Association for Computational Linguistics. Online: http://arXiv.org/abs/cs/0105030.

BIRD, STEVEN, and GARY SIMONS. 2003. Extending Dublin Core metadata to support the description and discovery of language resources. Computers and the Humanities 37, to appear.

BRADLEY, NEIL. 2002. The XML companion. Harlow, UK: Addison Wesley.

BRYLAWSKI, SAMUEL. 2002. Preservation of digitally recorded sound. Building a national strategy for preservation: Issues in digital media archiving. Washington, DC: Council on Library and Information Resources and the Library of Congress. Online: http://www.clir.org/pubs/reports/pub106/sound.html.

COHEN, ELIZABETH. 2001. Preservation of audio. Folk heritage collections in crisis. Washington, DC: Council on Library and Information Resources. Online: http://www.clir.org/pubs/reports/pub96/preservation.html.

COOMBS, JAMES H.; ALLEN H. RENEAR; and STEVEN J. DEROSE. 1987. Markup systems and the future of scholarly text processing. Communications of the ACM 30.933–47.

DCMI. 1999. Dublin Core Metadata Element Set, version 1.1: Reference description. Online: http://dublincore.org/documents/1999/07/02/dces/.

[63] http://www.opensource.org/

HARRISON, HELEN P. 1997. Audiovisual archives: A practical reader. Paris: UNESCO. Online: http://unesdoc.unesco.org/images/0010/001096/109612eo.pdf.

HIMMELMANN, NIKOLAUS P. 1998. Documentary and descriptive linguistics. Linguistics 36.161–95.

IANNELLA, RENATO. 2001. Digital rights management (DRM) architectures. D-Lib Magazine 7.6. Online: http://www.dlib.org/dlib/june01/iannella/06iannella.html.

INTERNATIONAL ASSOCIATION OF SOUND AND AUDIOVISUAL ARCHIVES. 2001. The safeguarding of the audio heritage: Ethics, principles and preservation strategy. Online: http://www.iasa-web.org/iasa0013.htm.

LADEFOGED, PETER. 2000. Vowels and consonants: An introduction to the sounds of languages. Cambridge, MA: Blackwell.

LAGOZE, CARL; HERBERT VAN DE SOMPEL; MICHAEL NELSON; and SIMEON WARNER. 2002. The Open Archives Initiative Protocol for Metadata Harvesting. Online: http://www.openarchives.org/OAI/openarchivesprotocol.html.

LAWLER, JOHN M., and HELEN ARISTAR DRY (eds.) 1998. Using computers in linguistics: A practical guide. London and New York: Routledge.

LEWIS, WILLIAM; SCOTT FARRAR; and D. TERENCE LANGENDOEN. 2001. Building a knowledge base of morphosyntactic terminology. Proceedings of the IRCS workshop on linguistic databases, ed. by Steven Bird, Peter Buneman, and Mark Liberman. Online: http://www.ldc.upenn.edu/annotation/database/ (Click on 'Papers').

LIBERMAN, MARK. 2000. Legal, ethical, and policy issues concerning the recording and publication of primary language materials. In Bird & Simons 2000.

LIBRARY OF CONGRESS. 1995. Guidelines for electronic preservation of visual materials. Online: http://lcweb.loc.gov/preserv/guide/.

LIBRARY OF CONGRESS. 2001. The deterioration and preservation of paper: Some essential facts. Online: http://lcweb.loc.gov/preserv/deterioratebrochure.html.

OCLC/RLG. 2002. Preservation metadata and the OAIS information model: A metadata framework to support the preservation of digital objects. Online: http://www.oclc.org/research/pmwg/pm_framework.pdf.

PULLUM, GEOFFREY K., and WILLIAM A. LADUSAW. 1986. Phonetic symbol guide. Chicago: University of Chicago Press.

SIMONS, GARY. 1998. The nature of linguistic data and the requirements of a computing environment for linguistic research. In Lawler & Aristar Dry 10–25. Appendix online: http://www.sil.org/computing/routledge/simons/.

SIMONS, GARY. 2000. Language identification in metadata descriptions of language archive holdings. In Bird & Simons 2000.

SIMONS, GARY, and STEVEN BIRD. 2002a. OLAC metadata. Online: http://www.language-archives.org/OLAC/metadata.html.

SIMONS, GARY, and STEVEN BIRD. 2002b. OLAC process. Online: http://www.language-archives.org/OLAC/process.html.

SIMONS, GARY, and STEVEN BIRD. 2002c. OLAC repositories. Online: http://www.language-archives.org/OLAC/repositories.html.

SIMONS, GARY, and STEVEN BIRD. 2003a. Building an Open Language Archives Community on the OAI foundation. Library Hi Tech 21.2.210–18. Online: http://www.arxiv.org/abs/cs.CL/0302021.

SIMONS, GARY, and STEVEN BIRD. 2003b. The Open Language Archives Community: An infrastructure for distributed archiving of language resources. Literary and Linguistic Computing 18.117–28.

ST-LAURENT, GILLES. 1996. The care and handling of recorded sound materials. Online: http://palimpsest.stanford.edu/byauth/st-laurent/care.html.

SYRACUSE UNIVERSITY LIBRARY. 1995. Procedures for recovering audio and sound recording materials. Online: http://libwww.syr.edu/information/preservation/audio.htm.

WACTLAR, HOWARD D., and MICHAEL G. CHRISTEL. 2002. Digital video archives: Managing through metadata. Building a national strategy for preservation: Issues in digital media archiving. Washington, DC: Council on Library and Information Resources and the Library of Congress. Online: http://www.clir.org/pubs/reports/pub106/video.html.

WALKER, JANICE R., and TODD TAYLOR. 1998. The Columbia guide to online style. New York: Columbia University Press. Companion site: http://www.columbia.edu/cu/cup/cgos/.

Wilkins, David. 1992. Linguistic research under Aboriginal control: A personal account of fieldwork in central Australia. Australian Journal of Linguistics 12.171–200.

Bird
Department of Computer Science
University of Melbourne
Victoria 3010
Australia
[sb@cs.mu.oz.au]

Simons
SIL International
7500 W. Camp Wisdom Rd.
Dallas, TX 75236
[gary_simons@sil.org]

[Received 23 July 2002;
accepted 27 March 2003]