From field data recording to online interlinear glossed text corpus

Daniel Kaufman
Queens College, CUNY & ELA
dkaufman@qc.cuny.edu

1 Introduction

1.1 Multipurpose documentation

- A Language Documentation mantra:
  - "a field of linguistic inquiry and practice in its own right which is concerned with the compilation and preservation of linguistic primary data and interfaces between primary data and various types of analyses based on these data" (Himmelmann 2006:1)
  - "a lasting, multipurpose record of a language" (Himmelmann 2006:1)
- We've shown how language documentation can be mobilized for a language community through popular channels. Here I'll focus on how linguists can utilize this documentation for descriptive purposes.

1.2 Transparency as a remedy to bad data

- The problem of bad data is a very serious one in linguistics. (Whole conferences have been dedicated to it!)
- How does "bad data" come about? Misunderstanding, confirmation bias, and arrogance on the part of linguists. Hyper-cooperation and misunderstanding on the part of the speakers.
- Examples are well documented in all fieldwork manuals (e.g. Bowern 2008; Chelliah and de Reuse 2011; Newman and Ratliff 2001; Vaux et al. 2007) yet very little has been done to enforce better practices or at least more transparency between claims and evidence.
- Every descriptive and theoretical claim should be backed up by primary data, not just another published assertion.
- Archiving is only one half of the solution: At a minimum, the data must be stored in a safe, stable and easily accessible location.
- But we also need to be able to search and browse that data in a convenient manner. Otherwise, all of our digital efforts are in vain.

1.3 Corpus linguistics

- A brief recent history of fieldwork based corpus linguistics (see E. Rafferty's concordance)
- Printouts > Shoebox > Toolbox > FLEx
- Corpus-aided/based linguistics vs. corpus-driven linguistics. Roughly:
  - **Corpus-aided**: Using the corpus to find examples (and counterexamples) of a linguistic phenomenon.
  - **Corpus-driven**: Using statistics generated by the corpus to tell us about variation, change and gradient phenomena (in addition to the above).
- Few corpora of endangered languages are large enough or well enough annotated to support *corpus-driven* linguistics.
  - For instance, our Wakhi corpus contains just under 70,000 words. Corpora for larger "global" languages contain many millions of words.
  - Corpora designed for syntactic research are fully parsed (cf. Tortora et al's recent AAPCAPPE[1] corpus). Every sentence has a corresponding tree structure or at least a representation of the argument structure. On the other hand, our Wakhi corpus is parsed morphologically but not syntactically.
- Despite these shortcomings, which make corpus-driven studies impossible, I will attempt to show how such corpora can be useful for research into grammatical systems.

1.4 FLEx & Kratylos

- FLEx is free, open source software developed by SIL for building linguistic databases consisting of a lexicon and interlinear glossed text (IGT).
- Kratylos[2] is a program being developed by Raphael Finkel and myself to share data generated in FLEx (among other programs) as an online corpus (supported by NSF DEL grant #1500753 and described in Kaufman and Finkel 2018).
- This aims to solve a serious problem in the documentation and archiving of endangered languages: Too few people are actually making use of the vast amount of archived documentation besides the linguists involved.
- At the very least, this requires a more viewing-friendly interface for non-linguists and a more research-friendly interface for linguists.
- As discussed in Kaufman and Finkel (2018), we have used popular available tools for the former population (e.g. YouTube, Facebook) and Kratylos for the latter population.

[1] https://aapcappe.commons.gc.cuny.edu/
[2] https://www.kratylos.org/

- The advantage of an IGT corpus over plain text is that the data is separated out by field (e.g. gloss, underlying form, allomorph, word category, etc.).
- The corpus also allows complex searches within and across specific fields using regular expressions.

Table 1: Regular expressions

`.` any character	`\b` word break	`\s` whitespace	`\S` non-whitespace
`*` zero or more times	`+` one or more times	`?` one or zero times	`{x}` *x* times
`{x,y}` at least *x* and at most *y* times		`^` beginning of line	`$` end of line
`[vcd]` *v*, *c* or *d*	`[^vcd]` not *v*, *c* or *d*	`dog|cat` *dog* or *cat*	
`(?!abc)` not followed by *abc*		`(?<!abc)` not preceded by *abc*	

- This can handle simple replacements such as `(NOM|OBL)` which matches with either NOM or OBL, but it can also handle far more complex contexts. The regular expression below matches a string containing A, not immediately followed by B but eventually followed by D without any intervening Cs: `(A) ?(?! ?(B)) (?!((?!D).)*?(C) )((\S+ ?)*?(D))`

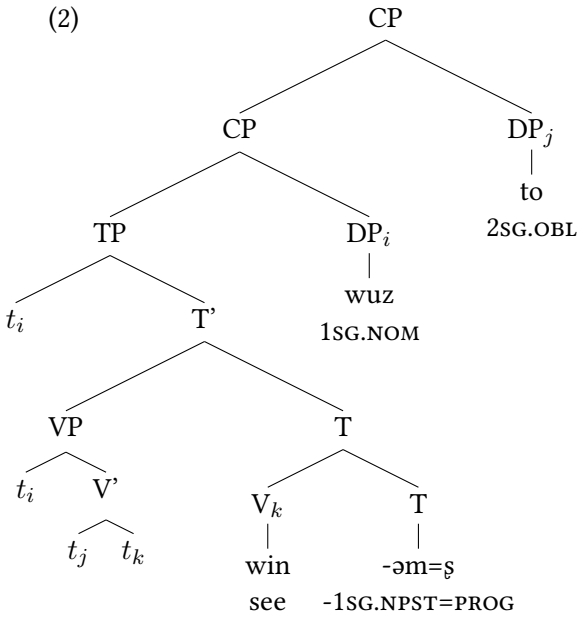
2 A simple test cases

2.1 Postverbal arguments in Wakhi

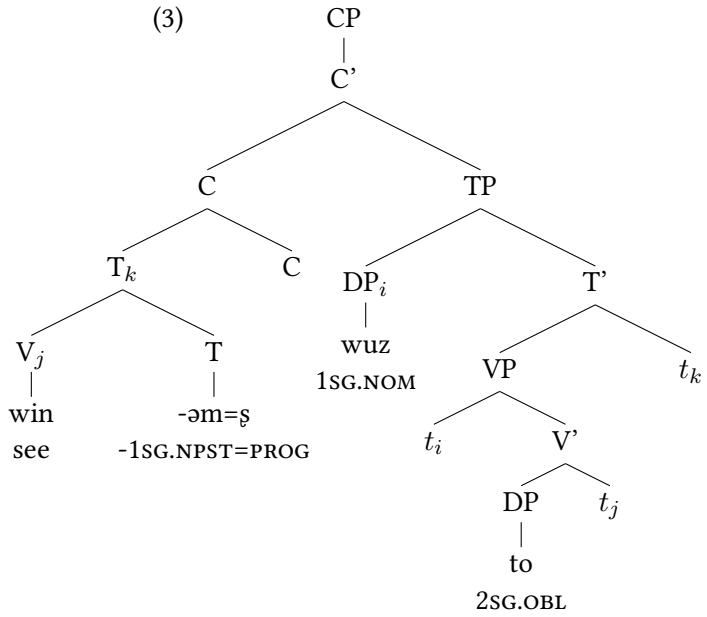
- Wakhi (Pakhalina 1975; Grünberg and Steblin-Kamensky 1988; Lorimer 1958) is an SOV language that allows for a good amount of scrambling, seemingly more than related languages like Persian/Tajik.
- Specifically, it allows for a post-verbal position of arguments where Persian/Tajik would demand a preverbal position.
- An elicited example is shown in (1). We want to know if this is derived by post-posing the subject and object, represented in (1), or by verb movement, represented in (2), among other options.

(1) win-əm=ṣ wuz **to**
see-1SG.NPST=PROG 1SG.NOM 2SG.OBL
'I see you.' (Columbia fieldmethods Fall 2011, 402)

(2)



(3)



- How can we do this? Displacing the arguments may correlate with a particular pragmatic status (e.g. topic, focus). Verb movement may correlate with a particular Mood (e.g. imperative, subjunctive, etc.).
- Searching for utterance-final occurrences of a nominative marked phrase with the regular expression `NOM$`, yields examples such as (4).
- Searching for `OBL$`, yields examples such as (5) and (6). What do these examples suggest?

(4) tu niv də rə-m ʧiz nung ʐɨ-nən=ət ɨkmət bɨrd podʃo **tu**
2SG.NOM now LOC ALL-PROX what name 1SG.GEN-ABL=and command win king 2SG.NOM
"Now, you here, governing and winning is my part, (but) you are the king."(ELA podʃo ət jaw kənd, WBL_2016_06_22a,16)

(5) agar ʐɨ soib pajdo vi-t-əj wuz taw-i nə=jund-əm agar
if 1SG.GEN owner apparent become-PST-PST 1SG.NOM 2SG.OBL-OBL NEG=take-1SG.NPST if
nə=vi-t-i jan jund-əm **taw-əj**
NEG=become-PST-PST then take-1SG.NPST 2SG.OBL-OBL
'...if my owner appears, I won't take you; if he doesn't appear then I will take you.' (GSK L'Oisillon Enchanté, 24.6)

(6) ðim-i kɨn-ən j-a-w-i wɨʐɨm-ən ʧuk-ən
prepared.clay-OBL dig-1PL/3PL.NPST DEM-MED-PRO-OBL bring-1PL/3PL.NPST strike-1PL/3PL.NPST
j-a-w-əj
DEM-MED-PRO-OBL
'We dig up the clay. We bring it back and we grind it up.' (GSK Augures lies a la cuisson du pain, 5.1)

- This pattern fits well with Frommer's (1981:135-145) corpus study of colloquial Persian, which shows a strong tendency for postposed arguments to be non-focal.
- We can further test this hypothesis by looking for postverbal interrogative expressions. As interrogative clauses are inherently focal, they should not be found in this position.
- The query `pst.*(what|who|where|when)$` seeks out all instances of an interrogative following a finite verb (one ending in either PST or NPST).

(7) potʃo j-a-w pərs-t-i ça-t-i **ʧiz**
king DEM-MED-PRO ask-PST-PST say-PST-PST what
"What is (the meaning of) this?' Asked the king.' (GSK tru nasiat, 17.4)

- But this is not a true counterexample. The interrogative is functioning here as a quote rather than a direct argument and quotes generally follow verbs of reporting in Wakhi.
- The other potential counterexamples turn out to be copular clauses, as in (8)-(10).

(8) çan-d xaj ti-n jan ti qsting gir **kuj**?
say-3SG.NPST okay 2SG.GEN-ABL then 2SG.GEN wrestling grasp who
'He said, 'Alright, yours then, who is your fighter?" (ELA diw qlajiʃ, WBL_2016_07_15, 1.41)

(9) jan çan-d ki tu=t **kuj**?
then say-3SG.NPST COMP 2SG.NOM=2SG who
'And she asked, 'Who are you?" (ELA Kɨngɨr, 1.123)

(10) j-a ðaj çan-d ki də-m ti bədʒəj **ʧiz**?
DEM-MED man say-3SG.NPST COMP LOC-PROX 2SG.GEN sack what
'The man asks: 'What have you got in this sack?" (PKH tsɨbɨr vrɨt, 1.7)

- When we dig further and look at nouns sandwiched between verbs and the end of the utterance with the query `v\S* n$`, we do find some striking differences with Tajik and Persian.
- Whereas in Tajik/Persian bare NPs are disallowed in postverbal position (Frommer 1981:142-145), Wakhi allows them there in one construction.

(11) awal dʒaj-i didʝ-ən xaʂ-ən d-r-a **ʝar**
first place-OBL see-1PL/3PL.NPST pull-1PL/3PL.NPST LOC-ALL-MED stone
'First we choose a spot and we bring over **stones**.' (GSK ʂəʝd xun çak, sɨj ʂkor, 2.1)

(12) tsə ʝar xʂa-ak-en kaʈ-ən **bənjod**
when stone pull-INF-ABL put-1PL/3PL.NPST foundation
'Once the stones have been brought over, we lay down **the foundation**.' (GSK ʂəʝd xun çak, sɨj ʂkor 2.2)

(13) was=əv ki kər-t-əj jan bər dʒoj tsar-ən **bijobon ʂung-v**
support_beam=3PL COMP do-PST-PST then on place do-1PL/3PL.NPST desert wood-OBL.PL
'When the main beam is set, we position the **roof beams** (desert wood).' (GSK ʂəjd xun çak - sij ʂkor, 2.8)

(14) bijobon ʂung-v=əv ki kər-t-əj kaʈ-ən **sparsk-v**
desert wood-OBL.PL=3PL COMP do-PST-PST put-1PL/3PL.NPST rafter-OBL.PL
'Once they do the roof beams, they put up the rafters.' (GSK ʂəjd xun çak - sij ʂkor, 2.9)

- The postverbal arguments here represent new information but seem to be most common when contrasted as part of a list, e.g. having done X, we do Y; having done Y, we do Z, etc.
- This suggests a focus hierarchy for the postverbal position across Iranian languages, as shown below. No language in this group seems to allow true interrogatives in postverbal position. This would be a big step towards SVO order.

Table 2: Focus hierarchy for postposed arguments

∅	>	OLD INFO	>	NEW INFO	>	INTERROGATIVES
Formal Persian		Colloq. Persian		Wakhi		

- Note that we've left the other hypothesis unexplored for now. The word order in (11)-(14) could be derived by verb movement. Note that the verbs in question ('put', 'do') are semantically light and function as light verbs. This is not borne out, however.

References

Bowern, Claire. 2008. *Linguistic Fieldwork: A Practical Guide*. New York: Palgrave Macmillan.

Chelliah, Shobhana L., and Willem J. de Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. Dordrecht: Springer.

Frommer, Paul Robert. 1981. Post-verbal phenomena in colloquial Persian syntax. PhD diss, University of Southern California.

Grünberg, Aleksander Leonovich, and Ivan M. Steblin-Kamensky. 1988. *La langue Wakhi*, volume 2: Essai grammatical et dictionnaire wakhi-francais. Paris: Maison des Sciences de l'Homme.

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for. In *Essentials of Language Documentation*, ed. Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1–30. Berlin and New York: Mouton de Gruyter.

Kaufman, Daniel, and Raphael Finkel. 2018. Kratylos: A tool for sharing interlinearized and lexical data in diverse formats. *Language Documentation & Conservation* 12:124–146.

Lorimer, David Lockhart Robertson. 1958. *The Wakhi Language*, volume I (Introduction, Phonetics and Texts). London: School of Oriental and African Studies, University of London.

Newman, Paul, and Martha Ratliff. 2001. *Linguistic Fieldwork*. Cambridge: Cambridge University Press.

Pakhalina, Tatyana N. 1975. *Vaxanskij jazyk*. Moscow: Nauka.

Tortora, Christina, Beatrice Santorini, Frances Blanchette, and C.E.A. Diertani. 2017. The Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE), version 0.1. URL: www.aapcappe.org

Vaux, Bert, Justin Cooper, and Emily Tucker. 2007. *Linguistic Field Methods*. Eugene: Wipf & Stock.