Informativity and affix order: a pilot study of Turkish

Sharon Inkelas[1] and Jem Orgun[2]

[1]University of California, Berkeley and [2]University of Colorado, Boulder

ABSTRACT

Informativity (contextual predictability) of a Turkish suffix in the context of the preceding morpheme is negatively correlated with distance from the root: on average, the more predictable in context, the farther from the root. This novel result contributes to the difficult task of predicting affix order in highly affixing languages, and extends the informativity metric to a new empirical domain.

PROBLEM

What grammatical properties predict affix order in highly affixing languages?

- Syntactic/semantic Scope (Baker 1988, Rice 2001)
- Semantic **relevance** (Bybee 1985)
- Phonotactics (P >> M; McCarthy & Prince 1994)
- Arbitrary templates; Simpson & Withgott 1986)
- A mixture of all of the above? (e.g. Rice 2011, Manova & Aronoff 2010, Saarinen and Hay 2014)

What usage-based properties have been shown to correlate with affix order?

- Relative base frequency (**parsability**); Hay 2002, Plag 2002, Hay & Plag 2004, Hay & Baayen 2005, Plag & Baayen 2009
- Bigram frequency (Ryan 2010)

HYPOTHESIS

Hypothesis: relative informativity (contextual predictability) correlates with relative distance from the root in words with more than one suffix in Turkish

Root-Suffix1-Suffix2-Suffix3….

If there is a correlation, do we expect it to be positive or negative?

INFORMATIVITY

- Measure of how predictable X is is context C.
- Negative log predictability of X in context C, weighted by predictability of C, averaged over all contexts in which X occurs
- Shown to correlate (inversely) with phone reduction (Cohen Priva 2013); the more informative (less contextually predictable) overall in a corpus, the less prone a given phone is to reduce generally
- See also Kuperman et al. 2007, Hirschberg & Pan 2000; Bell & Plag 2012a,b; Gahl & Garnsey 2004 on intonation and duration
- Within morphology, low entropy (high predictability) has been shown to facilitate learning (Moscoso del Prado Martín, Kostic & Baayen 2004, Blevins 2013; Ackerman & Malouf 2013; Seyfarth, Ackerman & Malouf 2014)

- Informativity is affected by some of the same factors that determine **parsability**, and by some of the same factors that influence **relevance**.
- Informativity is arguably an easier measure to apply than parsability, in that it doesn't necessarily require a huge corpus

THIS STUDY

This study: examines correlation with affix order, using Turkish as test case:

- S = some specific suffix in a given word.
- C = the immediately preceding morpheme (c ∈ C) (a bigram approach)

$$-\sum_{c} \frac{\Pr(C,S)\log_2 \Pr(S|C)}{\Pr(S)}$$

Summed over all contexts that the suffix occurs in

Weighted by predictability of that context

Negative log predictability of a suffix in a given context

TOY EXAMPLE

Lexicon:	-lar	'-pl'	Corpus:	[root]	200
	-a	'-dat'		[root]-a	50
				[root]-lar	100
				[root]-lar-a	10

Informativity of –lar:

$$= \frac{\Pr(\text{root}, \text{-lar}) * -\log_2 \Pr(\text{-lar}|\text{root})}{\Pr(\text{-lar})} = 2.55$$

Informativity of –a:

$$= \frac{\Pr(\text{root}, \text{-a}) * -\log_2 \Pr(\text{-a}|\text{root})}{\Pr(\text{-a})} + \frac{\Pr(\text{-lar}, \text{-a}) * -\log_2 \Pr(\text{-a}|\text{-lar})}{\Pr(\text{-a})} = 1.07$$

TURKISH CORPUS

- Original corpus: list of 611,390 uniquely parsed words from a large text corpus, each parsed into its constituent morphemes by Kemal Oflazer's morphological parser, in 2010 or so. Vowel harmony and other phonologically conditioned allomorphy is ignored in assigning affix IDs.
- The same word form may appear more than once in the corpus, in case it is ambiguous, but each parsed word is unique.

Examples

abartabilirler
(abart)abart +Verb+Pos(abil)^DB+Verb+Able(ir)+Aor(ler)+A3pl

abartabilirler
(abart)abart +Verb+Pos(abil)^DB+Verb+Able(ir)^DB+Adj+AorPart^DB +Verb+Zero+Pres(ler)+A3pl

abartabilirler
(abart)abart +Verb+Pos(abil)^DB+Verb+Able(ir)^DB+Adj+AorPart^DB +Noun+Zero(ler)+A3pl+Pnon+Nom

- Parsing tags were abbreviated and in some cases collapsed (obliterating overly sutble distinctions)
- The resulting file (word-glosses) has 531,919 lines = unique word parses.
- There are 78 unique affix types. (see table to right)

Three measures

- Suffix informativity (computed for each suffix over unique-word corpus)
- Suffix frequency (computed for each suffix over unique-word corpus)
- Suffix position index (computed for each suffix, as below)

➢ Divide words into 7 groups, defined by # of suffixes (range is from 2-8)
➢ Within each group, for each suffix S_i (of 78 total suffixes),
➢ Compute its average position, where root-adjacent = 0 and word-final = n), and divide by n
➢ The position index for a given suffix is the sum of these averages across all 7 word groups.

Position indices for suffixes in toy grammar, above

	Mean position, 1-sfx words	Mean position, 2-sfx words	position index
-a	0	1	1
-lar	0	.5	.5

-lar: closer to root (on average); more informative than *-a*
-a: farther from the root (on average); less informative than *-lar*

CORPUS EXAMPLE

otel-ci-ler-e
hotel-AGT-PL-DAT
'to the hoteliers'

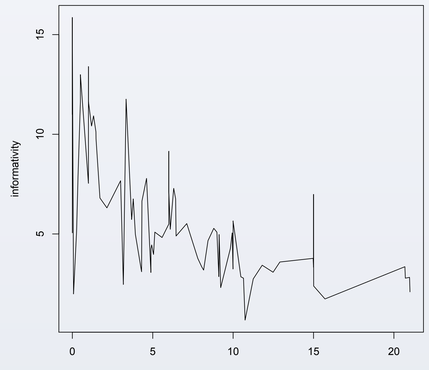
Observed suffix order in word	Corpus position index (increases)	(rank order)	Corpus Informativity (decreases)	(rank order)
1. AGT	3.69	23	5.73	46
2. PL	9.99	55	3.24	18
3. DAT	20.97	72	2.82	11

SUFFIX TABLE

suffix name	position index	(rank order)	informativity	(rank order)
ACQ	0.00	1	5.06	36
ADAMANTLY	0.00	2	15.86	73
RELATED	0.00	3	11.00	65
STAY	0.00	4	14.91	72
CAUS	0.07	5	1.99	3
BECOME	0.26	6	5.14	39
NEGAOR-NESS	0.35	7	7.88	59
DIM	0.50	8	11.64	66
REPEAT	0.50	9	12.99	70
ADVLY	1.00	10	7.54	56
IMP	1.00	11	12.12	69
INBETWEEN	1.00	12	13.39	71
SINCEDOING	1.00	13	11.65	67
EVERSINCE	1.19	14	10.42	63
ACTOF	1.32	15	10.92	64
FEELLIKE	1.47	16	10.12	62
HASTY	1.47	17	9.79	61
NML3	1.72	18	6.80	51
SANS	2.15	19	6.31	47
WITHOUT	3.00	20	7.67	57
PASS	3.17	21	2.47	7
UNABLE-NESS	3.34	22	11.76	68
AGT	3.69	23	5.73	46
ASIF	3.79	24	6.76	49
PLPOSS	3.92	25	4.98	34
NEGAOR	4.30	26	3.11	16
DESR	4.33	27	6.65	48
WHEN	4.61	28	7.79	58
INF1	4.88	29	3.07	14
UNABLE	4.89	30	4.27	27
PROG1	4.93	31	4.45	29
ABLE	5.05	32	3.97	26
FITFOR	5.13	33	5.09	37
SUBJREL	5.58	34	4.82	31
3SG	6.00	35	5.50	42
AFTERDOING	6.00	36	7.20	54
ASLONGAS	6.00	37	9.15	60
BYDOING	6.00	38	7.10	53
PASTPART	6.09	39	5.24	40
PROG2	6.30	40	7.29	55
NEC	6.42	41	6.76	50
NARRPART	6.45	42	4.90	32
INF2	7.12	43	5.52	43
COND	7.80	44	3.77	24
PASTPPL	8.04	45	3.37	21
AOR	8.16	46	3.19	17
2PL	8.44	47	4.67	30
NARR	8.80	48	5.28	41
JUSTLIKE	8.99	49	5.09	38
FUT	9.11	50	2.86	13
OPT	9.13	51	4.97	33
NESS	9.23	52	2.31	5
WITH	9.83	53	4.29	28
WHILE	9.94	54	5.05	35
PL	9.99	55	3.24	18
COPPL	10.00	56	5.66	45
SINCE	10.00	57	5.62	44
POSSPL	10.48	58	2.85	12
LOC	10.64	59	2.78	9
RELK11	10.74	60	0.68	1
3PL	11.25	61	2.75	8
PAST	11.81	62	3.43	22
1SG	12.49	63	3.08	15
1PL	12.91	64	3.59	23
INS	14.98	65	3.78	25
COP	15.00	66	3.34	19
EQ	15.00	67	6.99	52
2SG	15.01	68	2.39	6
POSS3	15.71	69	1.74	2
ABL	20.68	70	3.36	20
GEN	20.71	71	2.78	10
DAT	20.97	72	2.82	11
ACC	21.00	73	2.10	4

RESULTS

- The rank orders of position indices and informativity values for each suffix are negatively correlated (Spearman's rank correlation; rho = -.68, $p < .01$).
- ***Suffixes which tend to occur closer to the root are more informative (less predictable in local context) that those which tend to occur farther than the root***



Overall frequency is *not* correlated with position index or informativity

FUTURE STEPS

- Examine points of high discrepancy in the rank orders (e.g. Causative, Passive, Relativizer (= recursion sites? Sites governed strongly by scopal considerations?)
- Examine position/informativity behavior of the few suffixes that introduce recursion (e.g. Causative, Relativizer), vs. those that always occur in a fixed relative order with respect to one another
- Treat preceding string of morphemes in the word, not just immediately preceding suffix, as context (as Cohen Priva did for phones in his study of phone informativity)
- Build in word frequency data from large text corpus
- Individuate roots, rather than lumping into category 'ROOT'
- For homophonous affixes (ambiguous parses), estimate likelihood of one parse vs. another
- Apply method to another highly affixing language with a large tagged corpus, e.g. Finnish
- Compare Informativity to Relative Base Frequency (parsability) to see to what extent they converge or differ

REFERENCES

Ackerman, Farrell & Rob Malouf. 2013. Morphological organization: the low conditional entropy conjecture. *Language* 89(3). 429–464.
Baker, Mark. 1988. *Incorporation: a theory of grammatical function changing*. Chicago: Chicago University Press.
Balling, L. W. & Harald Baayen. 2012. Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition* 125(1). 80–106.
Bell, Melanie & Ingo Plag. 2012a. Compound stress, informativity and analogy. *Word Structure* 6(2). 129–155.
Bell, Melanie & Ingo Plag. 2012b. Informativeness is a determinant of compound stress in English. *Journal of Linguistics* 48. 485–520.
Blevins, James. 2013. The information-theoretic turn. *Psihologija* 46(3). 355–375.
Bybee, Joan. 1985. *Morphology: a study of the relation between meaning and form*. Amsterdam: Benjamins.
Gahl, Susanne & Susan Garnsey. 2004. Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80. 748–775.
Hay, Jennifer. 2002. From speech perception to morphology: Affix ordering revisited. *Language* 78(3).
Hay, Jennifer & Harald Baayen. 2005. Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences* 9.
Hay, Jennifer & Ingo Plag. 2004. What constrains possible suffix combinations? On the interaction of grammatical and processing restrictions in derivational morphology. *Natural Language and Linguistic Theory* 22.
Hirschberg, Julia & Shimei Pan. 2000. Modeling local context for speech accent prediction. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.
Kuperman, V, M Pluymaekers, M Ernestus & H Baayen. 2007. Morphological predictability and acoustic duration of interfixes in Dutch compounds. *Journal of the Acoustical Society of America* 12(4). 2261–2271.
Manova, Stella & Mark Aronoff. 2010. Modeling affix order. *Morphology* 20. 109–131.
Moscoso del Prado Martín, Fermín, Aleksandar Kostic & Harald Baayen. 2004. Putting the bits together: an information theoretical perspective on morphological processing. *Cognition* 94. 1–18.
Parker, Jeff & Andrea Sims. 2012. Affix ordering constraints and processing in Russian: a look at Complexity-Based ordering. University of Massachusetts, Amherst.
Plag, Ingo. 2002. The role of selectional restrictions, phonotactics and parsing in constraining suffix ordering in English. (Ed.) Geert Booij & Jaap van Marle. *Yearbook of Morphology 2001*. 285–314.
Plag, Ingo & Harald Baayen. 2009. Suffix ordering and morphological processing. *Language* 85(1).
Rice, Keren. 2000. *Morpheme order and semantic scope: word formation in the Athapaskan verb*. Cambridge: Cambridge University Press.
Rice, Keren. 2011. Principles of affix ordering: an overview. *Word Structure* 4. 169–200.
Saarinen, Pauliina & Jennifer Hay. 2014. Affix ordering in derivation. In Rochelle Lieber & Pavol Stekauer (eds.), *Oxford Handbook of Derivational Morphology*. Oxford: Oxford University Press.
Seyfarth, Scott, Farrell Ackerman & Rob Malouf. 2014. Implicative organization facilitates morphological learning. In Herman Leung, Zachary O'Hagan, Sarah Bakst, Auburn Lutzross, Jonathan Manker, Nicholas Rolle & Katie Sardinha (eds.), *Proceedings of BLS 40*, 480–494. Berkeley Linguistic Society.