35th Annual Boston University Conference on Language Development *Nov 6, 2010*

Cross-Linguistic Distributional Analyses with Frequent Frames: the Cases of German and Turkish

Hao Wang[1], Barbara Höhle[2], F. Nihan Ketrez[3], Aylin C. Küntay[4], Toben H. Mintz[1]

[1] University of Southern California, [2] University of Potsdam,
[3] İstanbul Bilgi University, [4]Koç University

Word Categorization

$$NP \rightarrow D \ \ N$$

- Syntactic categories as basic units of grammar
- Categorizing words is a necessary step for acquiring syntax

Current Approaches

- Phonological and prosodic cues
  - Cassidy & Kelly 1991, Kelly 1992, Monaghan, Christiansen & Chater 2007, Shi, Morgan & Allopenna 1998
- Semantic bootstrapping
  - Braine 1976, Grimshaw, 1981; Pinker, 1984, Schlesinger 1971
- Distributional information

Distributional information

- Lexical co-occurrence pattern
  - *the cat is on the mat*
- Originated from structural linguists
  - Bloomfield 1933, Harris 1951
- Advanced into a theory of language acquisition by Maratsos & Chalkley (1980)

Distributional Information in CDS

- Bigrams and other environments (Cartwright & Brent, 1997; Mintz, Newport & Bever, 2002; Redington, Chater & Finch, 1998)

  the __

- Frequent frames (Mintz 2003)
  - A frame is defined as two jointly occurring words with one word intervening.

  you __ the

Procedure for Frequent Frame Analysis

you read the story to Mommy

- *you read the*
- *read the story*
- *the story to*
- *story to Mommy*

you__the	**Verb frame**
the__to	**Noun frame**

would you put the cans back ?

you get the nuts .

you take the chair back .

you read the story to Mommy .

Categorization Evaluation

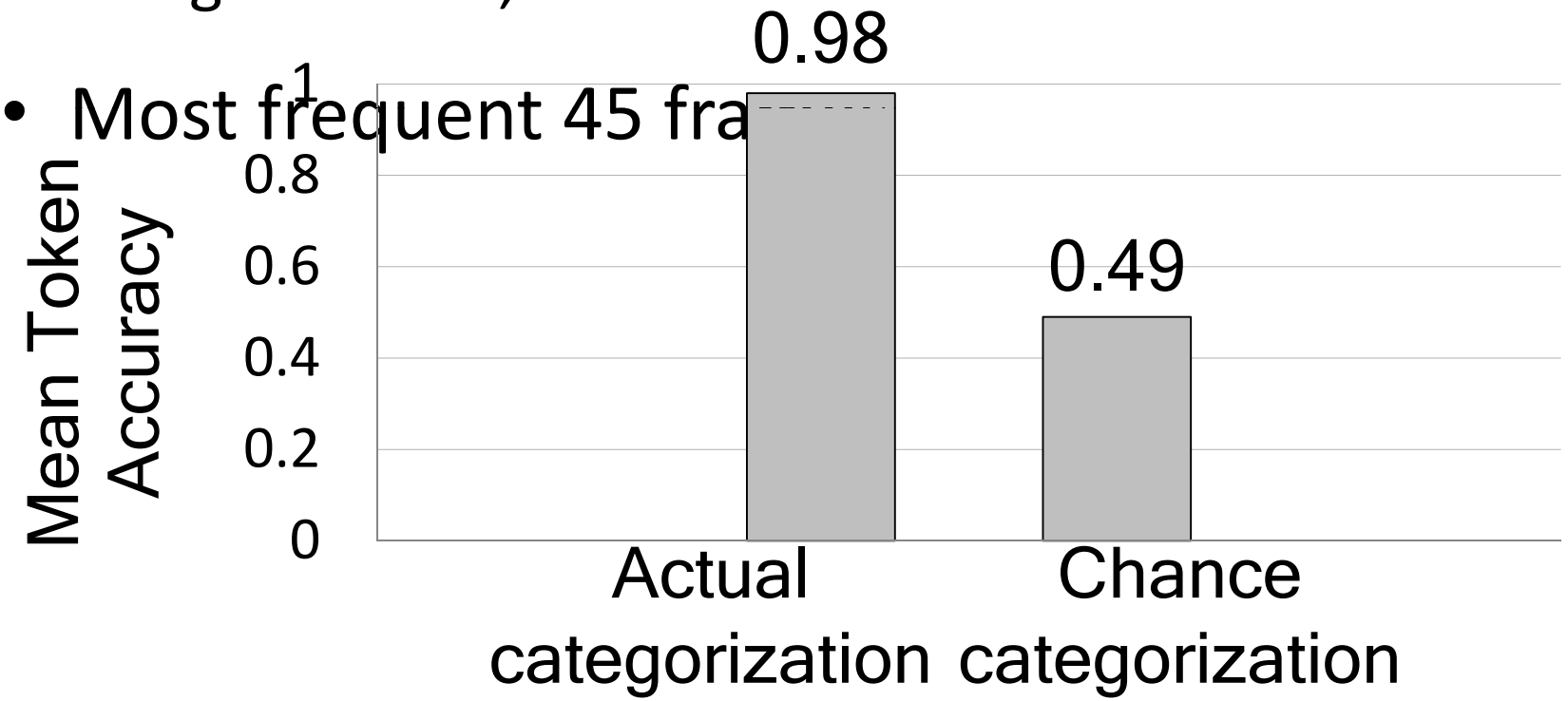
$$Accuracy = \frac{hits}{hits + false\_alarms}$$

- Accuracy is penalized when two words from different grammatical categories are grouped together
- Range: 0 to 1

Hit: two words from the same grammatical category grouped together
False alarm: two words from different grammatical categories grouped together

Frequent Frames in English CDS
(Experiment 1, Mintz, 2003)

- Six English corpora from CHILDES
  - Age under 2;6
- Most frequent 45 fra



you__it frame (Peter)

433 tokens **93 types**

put, see, do, did, want, fix, turned, get, got, turn, throw, closed, think, leave, take, open, find, bring, took, like, knocked, putting, pull, found, make, have, fixed, finish, try, swallow, opened, need, move, hold, give, fixing, drive, close, catch, threw, taking, screw, say, ride, pushing, hit, hiding, had, eat, carry, build, brought, write, wiping, wipe, wind, unzipped, underneath, turning, touching, tore, tie, tear, swallowed, squeeze, showing, show, said, rip, read, reach, pushed, push, play, pick, parking, made, love, left, knock, knew, hid, flush, finished, expected, dropped, drop, draw, covered, closing, call, broke, blow

Frequent frames in English

- Robust cue to word categories
- Potential source for initial bootstrapping
- Good accuracy even for a resource-limited learner (Wang & Mintz, 2008)

Learning problem

- Not only English-speaking children have to learn word categories. Every kid has to!
- Empirical questions
  - Work in typologically different languages?
  - Any modification to the mechanism?

Cross-linguistic Challenges

French	Homophony between clitic object pronouns and determiners. Three different definite determiners.
Spanish	Homophony among function words. Pro-drop.
Chinese	No inflectional morphology. Pro-drop.

Cross-linguistic studies

French	Chemla, Mintz, Bernal, & Christophe 2009 Most frequent 6 frames (at least 11 types, 18 tokens) **Accuracy = 1.0**, 12% corpus
Spanish	Weisleder & Waxman 2010 **Accuracy = 0.75**, 62% corpus
Chinese	Cai 2006, Xiao, Cai and Lee 2006 130 frames (at least 15 tokens) **Accuracy = 0.76**, Prominence = 90% (verb & adj.)
Dutch	Erkelens, 2009 **Accuracy = 0.71**

German Morphology

- A highly inflected language
  - 4 noun cases
  - 3 genders
  - 6 definite articles
- Many suppletive paradigms
  - der-das-die
  - er-ihn-ihm: he-him-him
- Determiners used pronominally

	'the man'
NOM:	der Mann
GEN:	des Mannes
DAT:	dem Mann
ACC:	den Mann

(Durrell, 2006)

Ein Journalist interviewte **die** Sängerin.
A jounalist interviewed the singer.

Die wurde auf der ganzen Welt als neuer Opernstar gefeiert.
She was celebrated across the whole world as the new opera star.

German Syntax

- Word order: Variable position of finite verb
  - Either 2nd or final position
  - All kinds of constituents can occur before the verb (subjects, objects, adverbials)

Er kauft Blumen
He buys flowers

weil er Blumen kauft
because he buys flowers

Turkish Morphology

- Agglutinative
  - Inflectional suffixes

ev	"house"
evler	"houses"
evlerim	"my houses"
evlerimiz	"our houses"
evlerimizde	"in our houses"
evlerimizdeki	"which is in our houses"

Turkish Syntax

- Free word order
  - SOV (canonical order)
    SVO, OVS, OSV, VSO, VOS
- Absence of function words
- Pro-drop

"Hakan went to school."
Hakan okula gitti.
Hakan gitti okula.
Okula Hakan gitti.
Okula gitti Hakan.
Gitti Hakan okula.
Gitti okula Hakan.

"Hakan **is** reading **the** book."
Hakan kitabı okuyor.
Hakan book-acc read-IPFV

"Hakan went to school."
Hakan okula gitti.
Ø okula gitti.
Hakan Ø gitti.

(Adapted from Underhill, 2006)

Challenges for Distributional Analysis

German	Flexible word order. Stem changes in def articles & pronouns. Determiners used pronominally.
Turkish	Free word order. Many bound morphemes. Absence of function words. Pro-drop.

Morpheme-Level Analysis

- Suffixes in languages with rich morphology
  - Many suffixes are restricted to be used with one word class
  - Turkish-learning children start producing inflections very early and they do it correctly
- Distributions at morpheme level

Current Analyses

- German
- Turkish
- Word-level frequent frames
- Morpheme-level frequent frames

Current Analyses – German CDS

- Simone (1;10.22-2;5.19) (Miller 1979)
  - 5685 utterances
- Most frequent 45 frames
- Control condition
  - Words randomly re-arranged between frames

soll ich mal gucken, was die machen ?
soll ich mal guck-en was die mach-en
soll ich mal guck -en was die mach -en

German Results

	Accuracy	Completeness	Tokens
FF Word	0.86	0.07	884
FF Morpheme	0.88	0.05	1857

German Frequent Morpheme Frames

Frame	Type	Token	Majority Cat.	%
was__-st	12	122	V	99%
Maxe__-t	32	107	V	100%
was__-t	18	91	V	100%
ge__-t	25	88	V	98%
-e__-e	32	65	Adj	38%
du__-st	26	65	V	98%
wir__-en	22	63	V	100%
'n__-chen	3	59	Pro	91%
-e__-en	21	57	V	82%
pass__auf	2	54	Pt	100%
-en__mal	11	52	Pro	44%
das__-t	18	49	V	100%

…

German Results

	Accuracy	Completeness	Tokens
FF Word	0.86	0.07	884
FF Morpheme	0.88	0.05	1857
_F_1F_2 Word	0.47	0.04	1216
F_1F_2_ Word	0.32	0.05	1462
_F_1F_2 Morpheme	0.78	0.10	2742
F_1F_2_ Morpheme	0.30	0.07	2672

Frequent Frames in German CDS (Stumper & Lieven, 2009)

- One child (2;0 – 2;6)
  - 28074 utterances
- Most frequent 45 frames

	Token accuracy		Type accuracy	
	Analysis Mean (SD)	Random Mean (SD)	Analysis Mean (SD)	Random Mean (SD)
Standard Labelling	.776[a,e,g] (.209)	.450[a] (.001)	.584[b,f,g] (.284)	.369[b] (.001)
Expanded Labelling	.643[c,e,h] (.198)	.335[c] (.001)	.432[d,f,h] (.213)	.257[d] (.001)

[a,b,c,d] Scores differ significantly (Fisher's Omnibus Test, $p < .001$).
[e,f,g,h] Means differ significantly (paired t-tests, $p < .001$).

Current Analyses – Turkish CDS

- Two children (Ural, Yuret, Ketrez, Koçbaş & Küntay 2009)
  - Elif (0;9.10-1;9.28), 21741 utterances
  - Irmak (0;9.0-2;0.16), 16024 utterances
- Control condition
  - Words randomly re-arranged between frames

sen hep ayaklarını sokuyo(r)sun .
sen hep ayak-lar-ın-ı sok-uyo(r)-sun .
sen hep ayak-PL-POSS&2S-ACC sok-IPFV-2S
sen hep ayak PL POSS&2S ACC sok IPFV 2S

Turkish Results

	Corpus	Accuracy	Completeness	Tokens
FF Word	Elif	0.54	0.09	1269
	Irmak	0.40	0.11	1656
FF Morpheme	Elif	0.93	0.06	6102
	Irmak	0.88	0.06	2764

Turkish Frequent Morpheme Frames (Elif)

Frame	Type	Token	Majority Cat.	%
GEN__POSS&3S	163	538	N	96%
ne__IPFV	24	348	V	100%
ne__PAST	26	316	V	98%
QUE__PAST	77	260	V	98%
DAT__PAST	43	217	V	100%
QUE__IPFV	59	215	V	99%
DAT__IPFV	51	209	V	100%
ACC__PAST	71	203	V	100%
QUE__FUT	34	165	V	100%
LOC__var	52	152	WH	55%
ACC__IPFV	61	151	V	99%
…				

Turkish Results

		Corpus	Accuracy	Completeness	Tokens
FF Word		Elif	0.54	0.09	1269
		Irmak	0.40	0.11	1656
FF Morpheme		Elif	0.93	0.06	6102
		Irmak	0.88	0.06	2764
Bigram Morpheme	F_1_	Elif	0.31	0.05	33793
		Irmak	0.39	0.05	16678
	_F_1	Elif	0.66	0.10	41540
		Irmak	0.72	0.09	29017

Conclusion

- Frequent morpheme frames are highly accurate categorization contexts in German and Turkish
- Languages with rich morphology and free word order are not a problem for frequent frames, if analyzed at the right level

A plausible mechanism

- Young children have access to morpheme-level information
  - Aksu-Koç & Ketrez 2003, Ketrez & Aksu-Koç 2009, Hohle, Schmitz, Santelmann & Weissenborn 2006, Santelmann & Jusczyk 1998
- Dutch infants can use frequent morpheme frames to categorize nonsense words (Erkelens, 2009)
- Building syntactic representation from morpheme-level distributions

Bootstrapping to word categories

- Target language dependent distributional analysis
  - Determine granularity of analysis from statistical regularities (Saffran, Aslin & Newport 1996, Swingley 2005)
- Integration with other sources of information
  - Phonological cues, semantics, language specific distributional cues
- Frequent frames as a potential universal cue for initial bootstrapping

Acknowledgements

- Dilara Koçbaş for transcribing and coding the Turkish data
- Frauke Berger for labeling German data
- USC Psychology Developmental Brown bag

This research was supported by Turkish Academy of Sciences, in the framework of the Young Scientist Award Program granted to Aylin C. Küntay (EA-TÜBA-GEBİP/2001-2-13) and supported in part by a grant from the National Science Foundation (BCS-0721328) to Toben H. Mintz.

Category Labels in German

Adj	Adv	Conj	Det
Interj	N	Prep	Pro
Pt	V	WH	

Category Labels in Turkish

ADJ	ADV	ART	CO
CONJ	EXIST	INTERJ	N
NEG	NUM	POST	PRO
V	WH		

Turkish Morpheme Labels

1P	1st person plural
1S	1st person singular
2S	2nd person singular
ABIL	Abilitative mood
ABL	Ablative case
ACC	Accusative case
AOR	Aorist (present tense marker)
CAUS	causative
CM	compound marker
DAT	dative case
DIM	diminutive
FUT	future tense
GEN	genitive case
INF	infinitive
IPFV	imperfective (progressive marker)
LOC	locative case
NEG	negation
OPT&1P	optative and 1st person plural fused
OPT&1S	optative and 1st persin singular fused
OPT&3S	optative and 3rd persin singular fused
PAST	past tense
PFV	perfective aspect
PL	plural
POSS&1S	possessive and 1st persin singular fused
POSS&2S	possessive and 2nd person singular fused
POSS&3S	possessive and 3rd person singular fused
QUE	yes-no question particle

German Results

	Actual			Random	
	Accuracy	Completeness	Tokens	Accuracy	Completeness
FF Word	0.86	0.07	884	0.37	0.03
FF Mor	0.88	0.05	1857	0.51	0.03
_F_1F_2 Word	0.47	0.04	1216	0.31	0.03
F_1F_2_ Word	0.32	0.05	1462	0.17	0.03
_F_1F_2 Mor	0.78	0.10	2742	0.32	0.04
F_1F_2_ Mor	0.30	0.07	2672	0.16	0.03

Turkish Results

		Actual			Random	
	Corpus	Accuracy	Completeness	Tokens	Accuracy	Completeness
FF Word	Elif	0.54	0.09	1269	0.19	0.03
	Irmak	0.40	0.11	1656	0.27	0.08
FF Mor	Elif	0.93	0.06	6102	0.49	0.03
	Irmak	0.88	0.06	2764	0.38	0.03
Bigram F_1	Elif	0.31	0.05	33793	0.20	0.03
	Irmak	0.39	0.05	16678	0.26	0.04
Mor F_1	Irmak	0.66	0.10	41540	0.24	0.04
	Irmak	0.72	0.09	29017	0.29	0.03

German Frequent Frames

Frame	Type	Token	Majority Cat.	%
was__denn	19	56	V	100%
pass__auf	2	54	Pt	100%
was__'n	10	53	V	100%
wir__mal	20	37	V	81%
wo__denn	7	30	V	100%
was__Mone	6	24	V	87%
komm__her	1	23	Pt	100%
was__der	5	22	V	100%
ist__der	9	22	Pt	50%
ist__das	3	22	Pt	95%
Maxe__dir	8	21	V	100%
…				

German Morpheme Frame
Maxe__-t

107 tokens 32 types

- mach, ha, gib, trink, zeig, zieh, sag, hol, guck, spritz, hilf, komm, brauch, nimm, setz, faell, tu, fang, krieg, sieh, schneide, bring, spiel, schmier, schlaef, iss, stoess, wisch, träg, schieb, hael, blaes

Turkish Frequent Frames

Frame	Type	Token	Majority Cat.	%
gel__gel	29	69	V	34%
burada__var	23	68	WH	72%
ne__Ekin	32	58	V	79%
orada__var	34	55	N	52%
ne__orada	7	53	EXIST	83%
bir__daha	12	47	N	100%
çok__bir	24	41	ADJ	80%
çok__değilmi	21	39	ADJ	69%
Ekin__Ekin	18	34	WH	47%
da__var	29	34	N	88%
bak__bak	23	33	N	42%

…

Turkish Morpheme Frame
ACC__PAST

203 tokens **71 types**

- gör, bul, at, yap, ver, al, unut, ye, çiz, boya, göster, duy, boz, soy, tak, yakala, koy, kandır, yık, çarp, dön, aç, çağır, beğen, iç, ara, yıka, öğren, bit, kaldır, dağıt, ısır, acı, yol, ez, korkut, mahvet, üz, giy, kır, dinle, götür, yala, devir, süsle, sok, yut, sula, gıdıkla, kazan, oy, dik, ek, tamamla, sür, topla, çıkar, kes, parçala, as, seyret, kaydet, de, çal, çık, git, yırt, kaşı, söyle, özle, şaşır

English Frequent Frames (Eve)

Frame	Type	Token	Major Cat.	%
what__you	9	353	aux	99%
you__to	20	235	v	70%
you__it	72	207	v	99%
you__the	46	119	v	95%
you__a	27	117	v	93%
are__doing	3	110	pro	100%
what__that	6	108	cop	93%
you__me	16	104	v	100%
would__like	4	96	pro	97%
to__it	37	88	v	98%
you__have	16	87	aux	94%
you__your	31	87	v	95%

…