Particle placement in EFL learner speech: core probabilistic grammar and/or EFL-specific preferences?

Magali Paquot

Centre for English Corpus Linguistics
UCLouvain

Introduction

- « Exploring probabilistic grammar(s) in varieties of English around the world » (Project director: Benedikt Szmrecsanyi; FWO, grant # G.0C59.13N)
- Overarching objective
  - « understand the plasticity of the probabilistic knowledge of English grammar on the part of language users with diverse regional and cultural backgrounds » (Szmrecsanyi et al., to appear)

« Exploring probabilistic grammar(s) in varieties of English »

- Investigate the extent to which …
  - Varieties of English share a core probabilistic grammar
    - the choice between syntactic alternations is motivated by probabilistic constraints rather than categorical rules (cf. Bresnan, 2007).
  - Grammatical variation is also subject to indigenization

Szmrecsanyi et al. (to appear)

« Exploring probabilistic grammar(s) in varieties of English »

- Three syntactic alternations
  - Particle placement, genitive and dative alternations
- Up to nine varieties of English as represented in the *International Corpus of English* and the *GLoWbE*
  - British, Canadian, Hong-Kong, Indian, Irish, Jamaican, New Zealand, Philippine and Singapore English
- Focus on users of first and second language varieties of English

Szmrecsanyi et al. (to appear)

KULeuven – UCLouvain collaboration

- Shed some light on whether **English as a Foreign Language (EFL) learners** share a core probabilistic grammar with users of first and second language varieties of English.

- Started in December 2015
  - Work very MUCH in progress …
  - Student worker (Sarah Pesenti) for a full month at UCLouvain
    - Thank you very much again!

Particle placement alternation

- Exhibits the most robust variety effects (Szmrecsanyi et al., to appear)
  - Variety is ranked as the single most important predictor of particle placement choice by conditional random forest analysis

Phrasal verbs and EFL

- Notoriously difficult for EFL learners (e.g. Celce-Murcia & Larsen-Freeman, 1999)
- Avoidance / underuse
- Role of the L1
  - L1 Bulgarian, Dutch, German, Russian and Swedish learners show less avoidance than L1 French, Hebrew, Italian and Spanish learners (Alejo González, 2010; Dagut & Laufer, 1985; Gilquin, 2015; Hulstijn & Marchena, 1989; Sjöholm, 1995; Waibel, 2008)
- Errors, deviations, unnaturalness
  - *carry out a race, make up a proposal* (Waibel, 2008)
  - *find back, see his mother back* (Gilquin, 2015)

Research questions

- What factors influence EFL learners' particle placement alternation?
- How do EFL learners' particle placement preferences compare with those of users of first and second language varieties of English?
  - Cf. Szmrecsanyi et al. (to appear)

Data

- Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al. 2010)
  - All the components follow the same structure, with c. 50 interviews made up of three tasks: set topic, free discussion and picture description.
  - 11 mother tongue backgrounds (Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish and Swedish)
  - Intermediate learners
- Louvain Corpus of Native English Conversation (LOCNEC; De Cock 2004)
  - a comparable corpus of interviews with native speakers of English

LINDSEI-FR transcript

- <B> (er) about six years ago I went with my family to Sicily . (er) because my parents were tired of . of their jobs and they wanted to escape a bit </B>
- <A> <overlap /> have a big break </A>
- <B> <overlap /> a little bit yes have have a big break and during the[i:] Easter break we: we (er) went to . there and (erm) also because my my father is a teacher in classical languages and so he is . always interested in (er) Greek Antiquity and . and and and so on and so it was particularly interesting for him . and (em) the country is really (eh) beautiful but (er) quite poor .. so (eh) we went to the: (em) .. ty= typical (er) . areas so where where there are (eh) Roman temples </B>
- <A> <overlap /> (mhm) </A>
- <B> <overlap /> Greek temples in ruins to (er) Roman theatres et cetera and (er) we also went to: (eh) villages (er) and (er) most of the time villages are really poor and dirty and (er) it's completely different from here .. (er) and (eh) people there are really (em) . terrific when when they are driving and they do not stop at (eh) red lights they <overlap /> <XX> </B>

	Interviews	Words (learners only)
LINDSEI-FR	50	94,941
LINDSEI-GE	50	89,384
LINDSEI-SW	50	75,202
LOCNEC	50	125,069

Method

- Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller & Melanie Röthlisberger (to appear). "Around the world in three alternations: modeling syntactic variation in varieties of English". English World-Wide 37(2).
- *Annotation of common features for the genitive, dative, and particle placement alternations* (Grafmiller et al, 2015)
- *Guidelines for selection and annotation of interchangeable particle verbs* (Grafmiller, 2015)

Data selection and extraction

- Interchangeable transitive particle verbs
  - *around, away, back, down, in, off, out, over, on, up*
- Raw corpus data
  - Relatively small learner corpora
  - Hesitations, disfluencies, repeats, etc.
- Manual weeding-out
  - Prepositions, adverbs, etc.
  - Prepositional verbs
  - Tokens that did not include genuinely interchangeable uses
    - Passive sentences, sentences with extracted direct objects, modified particles, fixed phrases, etc.

	LINDSEI-FR	LINDSEI-GE	LINDSEI-SW	LOCNEC
around	35	60	47	72
away	23	28	29	88
back	94	82	73	226
down	10	32	33	127
in	1773	1399	*1077*	*2180*
off	14	24	30	133
on	441	378	*290*	*874*
out	54	93	104	304
over	16	68	42	122
up	45	101	117	318
Total	**2505**	**2265**	**1842**	**4444**

Transitive particle verbs

	LINDSEI-FR	LINDSEI-GE	LINDSEI-SW	LOCNEC
V-P-DO	13 (35.1%)	34 (50.7%)	27 (39.7%)	39 (24.2%)
V-DO-P	24 (64.9%)	33 (49.3%)	41 (60.3%)	122 (75.8%)
Total	**37**	**67**	**68**	**161**
	(1.47%)	(2.96%)	*(3.69%)*	*(3.62%)*

333

Predictor variables (1)

- VARIETY
  - FR, GE, SW, EN
- NATIVE
  - L2, L1
- DIROBJTYPE
  - nc: common noun
  - np: proper noun
  - pprn: personal pronoun
  - iprn: impersonal pronoun
  - dm: demonstrative pronoun
  - ng: gerund
- DIROBJWORDLENGHTH
  - Length in words of the direct object

Predictor variables (2)

- DIROBJLETTERLTH
- DIROBJANIMACY (Wolk et al. 2013)
  - a: human & animal
  - c: collective
  - i: inanimate
  - l: locative
  - t: temporal
- DIROBJDEFINITENESS (Garretson et al., 2004)
  - def: proper nouns, NPs with definite determiner, definite pronouns, s-genitive NPs, superlatives, temporal expressions
  - Indef: NPs with indefinite determiner, indefinite pronouns, bare plural NPs, …

Predictor variables (3): work in progress

- DIROBJGIVENNESS
  - whether a noun had been mentioned recently in the discourse: a constituent was coded as ‘given’ if its head noun (lemma) was mentioned in the 100 words prior to the actual occurrence, and as ‘new’ otherwise.
- DIROBJTHEMATICITY
  - the extent to which a word represents the topic or "theme" of a text
  - relative frequency of a head noun in the text in which it occurs.
- DIROBJHEADFREQ (pmw)
  - *British National Corpus*

Predictor variables (4): work in progress

- PPDIRECTIONAL
  - The presence of a directional PP following the target VP
- VERBSEMANTICS (Gries, 2003)
  - Literal
  - Metaphorical
  - Idiomatic

Modelling

- LINDSEI-FR, LINDSEI-GE, LINDSEI-SW
  - Resp ~ Variety + DirObjWordLth + DirObjLetterLth + DirObjType + DirObjAnimacy + DirObjGivenness + DirObjThematicity + DirObjHeadFreq + DirObjDefiniteness + DirectionalPP + VerbSemantics

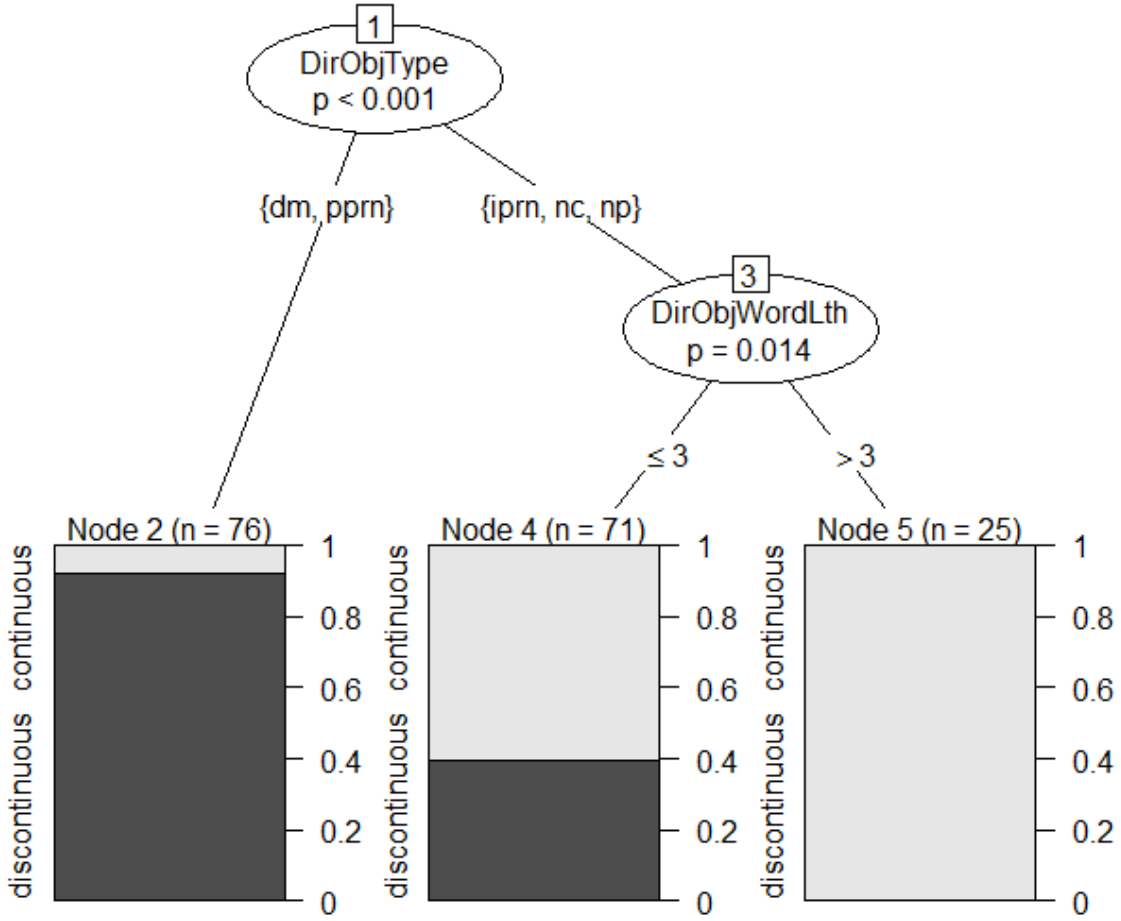
- LINDSEI-FR, LINDSEI-GE, LINDSEI-SW, LOCNEC
  - Resp ~ native + Variety + DirObjWordLth + DirObjType + DirObjAnimacy + DirObjDefiniteness

Effects of variables

- Conditional inference trees
  - predict outcomes by recursively partitioning the data into smaller and smaller subsets according to those predictors that co-vary most strongly with the outcome
  - Visualization of interactions among predictors
- Conditional Random Forest
  - Measure of the overall importance of each predictor
- R (R Core Team, 2015); *party* package; *Hmisc* package

Tagliamonte & Baayen (2012), Szmrecsanyi et al. (to appear)

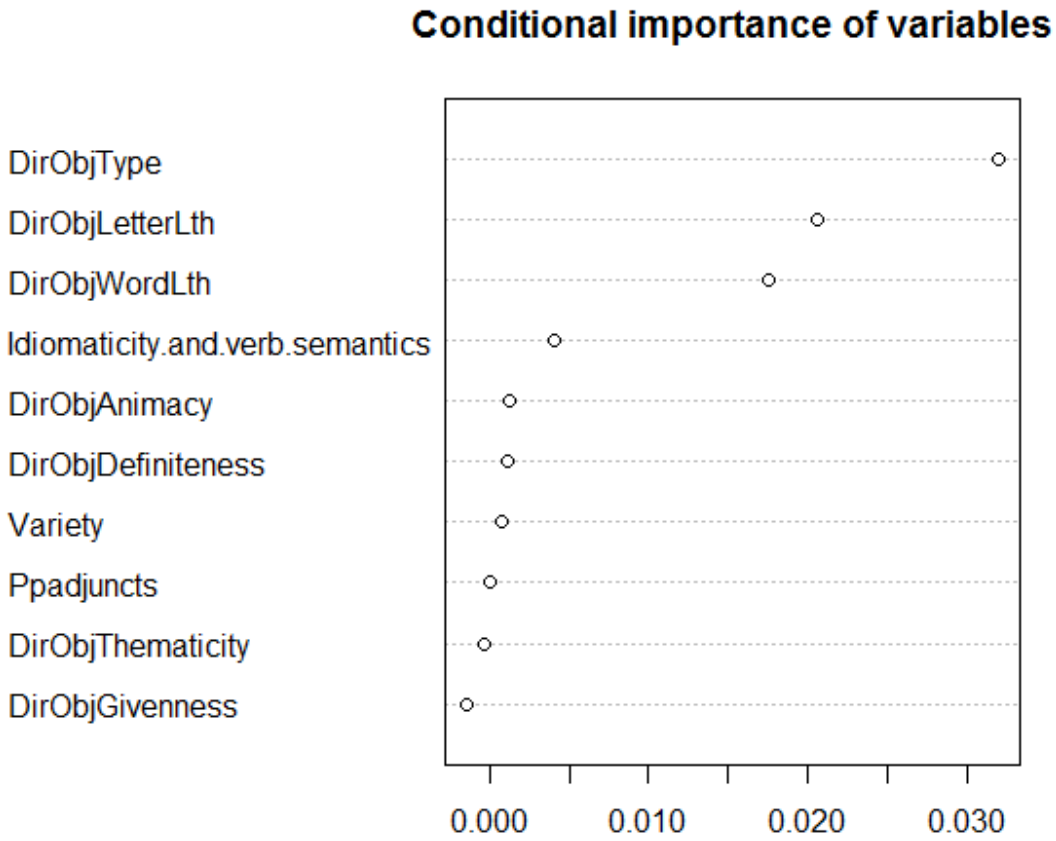
PRELIMINARY RESULTS



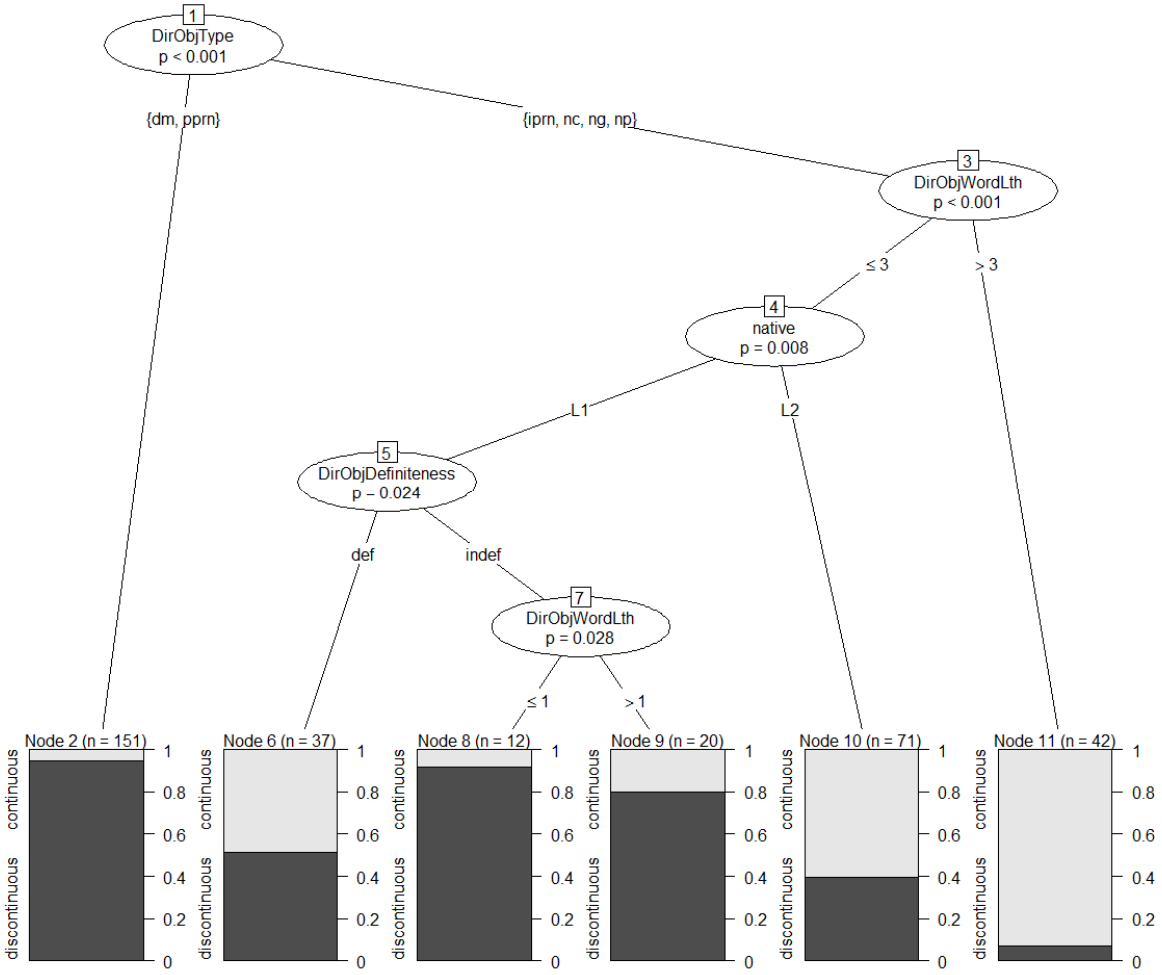
Learner data only [C = 0.86]

lindsei.ct <- ctree(Resp ~ Variety + DirObjWordLth + DirObjLetterLth + DirObjType + DirObjAnimacy + DirObjGivenness + DirObjThematicity + DirObjHeadFreq + DirObjDefiniteness + DirectionalPP + Idiomaticity.and.verb.semantics)

LINDSEI-FR, LINDSEI-GE & LINDSEI-SW



C = 0.94



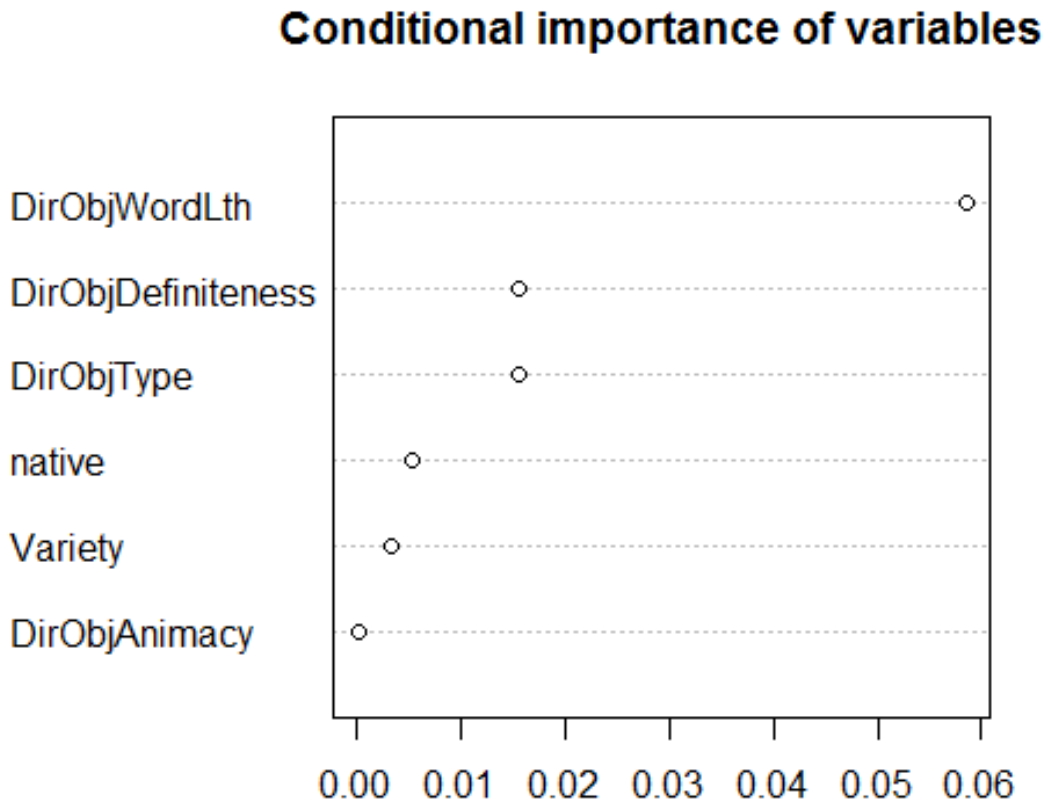
LINDSEI + LOCNEC [C = 0.88]

lindsei_locnec2.ct <- ctree(Resp ~ native + Variety + DirObjWordLth + DirObjType + DirObjAnimacy + DirObjDefiniteness)

Indefinite NPs in V-DO-P

- *take a year off / out* (very frequent)
  - *I might . I might take a year out in France*
  - *cos I took a year out before I came here*
  - *through and I'd already taken a year off so I decided just*
- *sort things out, pick things up*
  - Bare plural NPs
  - Respective weight of different variables
    - Simple objects
    - Direct objects without a determiner result in a significant preference for V-DO-P (Gries, 2003: 86)

LINDSEI-FR, LINDSEI-GE, LINDSEI-SW & LOCNEC



C = 0.91

Very preliminary (!) answers to RQs

- What factors influence EFL learners' particle placement alternation?
  - Type of direct object
  - Length of the direct object (number of words)
  - (Verb semantics)
- How do EFL learners' particle placement preferences compare with those of users of first and second language varieties of English?
  - Bias towards V-Part-DO (// L2 varieties)
  - Not sensitive to other factors that proved influential in the literature ?
    - Not sensitive to (in)definiteness // first language acquisition (Gries, 2011)

Core probabilistic grammar and/or EFL-specific preferences?

- Core probabilistic grammar
  - L1-independent generalizations
  - Effect directions of the variables (length, object type, verb semantics) are stable across L1s
- EFL-specific preferences
  - Clear % L1 English / EFL
  - No effect of the L1
    - Avoidance/underuse

Limitation

- Size of dataset!
  - Number of cases per ‘variety’
  - Number of EFL ‘varieties’

In the pipeline (1)

- Finish off LOCNEC + LINDSEI-SW analyses
- Analyze more LINDSEI components

- Writing data
  - *International Corpus of Learner English* (Granger et al., 2009)

- Semi-automatic procedure?
  - POS tagging: precision and recall
  - Regular expressions

- Student internship: Marie Gabrys

In the pipeline (2)

- Other (external) variables
  - Frequency of PV
    - Higher phrasal verb frequency is correlated with (higher percentages of) V-DO-Prt (Gries, 2011)
  - Association between verb & particle (e.g. Delta P)
  - Association between PV and V-P-DO/V-DO-P (Gries & Stefanowitsch, 2004)

Table 4. Distinctive collexemes for *[V Prt Obj]* and *[V Obj Prt]*

V Prt Obj (N=1,251)		V Obj Prt (N=1,192)	
Collexeme	*Distinctiveness*	*Collexeme*	*Distinctiveness*
carry out (49:1)	9.10E-14	get back (0:18)	2.30E-06
find out (49:5)	3.83E-10	get out (2:21)	1.91E-05
point out (43:3)	4.42E-10	play back (1:12)	0.0013
set up (42:8)	1.06E-06	turn off (2:14)	0.0015
take on (37:7)	4.60E-06	ring up (3:16)	0.0015

In the pipeline (3)

- Mixed-effect modeling
  - Random effects: speaker, verb, particle, object head

- Other alternations
  - MA dissertation on genitive alternation (Sarah Pesenti)

Thank you to …

- Benedikt Szmrecsanyi
- Jason Grafmiller
- Sarah Pesenti (student worker)

Thank you for your attention!

Questions? Comments? Suggestions?