Modeling reduction of *is, am* and *are* in grammaticalized constructions

Danielle Barth
University of Oregon

March 29, 2011

Quantitative Investigations in Theoretical Linguistics 4

- Background Information

  Grammaticalization

  Gramaticalization and Reduction

  Frequency and Reduction

- The Case of *is, am* and *are*
- The Corpus
- Type of Statistical Model
- Results
- Discussion

Grammaticalization

- A type of language change

- creation of grammatical element from a lexical element or another grammatical element

  ex: English *will* ‘want’ > *will* FUTURE

- sometimes accompanied by phonological reduction of the grammaticalized word

  ex: English *I’**ll** see you later*

  but not: **I’**ll** it to be so*

Grammaticalization

- results in new paradigmatic and syntagmatic uses and limitations

- sometimes results in a change of form
  - a reduction in length
  - loss of vowels
  - devoicing
  - loss of final consonants

Grammaticalization and Reduction

- Why do grammaticalized elements reduce?
  - low semantic weight (Bybee and Pagliuca 1985, Gabelentz 1891, Givon 1985, Heine 1993, Hopper and Traugott 1993, Lehmann 1995)
  - frequency of use (Bybee 2007)
  - separate storage in mental lexicon, as homonyms, is required for both these explanations

Frequency and Reduction

- Why do frequent elements reduce?
  - expected words are produced faster and less clearly than surprising words (Pierrehumbert 2002)
  - listeners build up memories of hypo-articulated forms of frequent words, and then in turn use these memories to produce their own speech, further entrenching the idea of a lenition-bias on frequent forms (Pierrehumbert 2001, 2002)

Frequency and Reduction

- Lexical words: Homonyms with different frequencies have different lengths and more frequent words are shorter (Gahl 2008)

- Grammatical words: frequency is an explanatory factor for reduced vowel production in the most frequent meanings of *that* and *of* (Bell et al. 2003)

Frequency and Reduction

- Lexical v. grammatical morphemes: grammatical morphemes are shorter than their lexical homophonous morphemes in Dutch (van Bergem 1995)

- For highly frequent function words and their content word homophones, following conditional probability (P(A|B)) predicted reduction (Bell et al 2009)

Reduction

- There are lots of other reasons for phonological/phonetic reduction aside from grammaticalization (Bybee 2007, van Bergem 1995)

- Could theoretically have a case where the source construction reduces and the grammaticalized construction doesn’t reduce

The case of *is, am* and *are*

- Grammaticalization research tells us that the grammaticalized, more grammatical variant is supposed to reduce in relation to its source construction, due to a decrease in semantic weight

- Frequency research tells us that the more frequent homonym will reduce more than a less frequent homonym

The case of *is, am* and *are*

- English *be* in the copula construction is the source for the grammaticalized progressive and passive constructions
- In this study, inflections of *be* investigated are *is, am* and *are*
- Both the source and grammaticalized elements can reduce

She is a welder	*She's a welder*
She is working	*She's working*
She is seen	*She's seen*

The case of *is, am* and *are*

- The source copular construction is also semantically empty
- The source copular construction is much more frequent than either of the grammaticalized constructions

	's	*is*	*are*	*'re*	*'m*	*am*	**Total**
Copula	611,889	579,515	205,514	96,982	89,619	11,711	1,586,230
Progressive	97,627	110,017	105,696	164,067	55,338	3,426	536,171
Passive	43,137	54,190	40,736	16,657	5,097	1,300	161,117
Total	752,653	743,722	351,946	277,706	150,054	16,437	2,292,518

COCA totals for Tokens of Interest by Construction Type as of Nov 19, 2010

Historical Summary

- In Old English
  - The copula construction
  - The forerunner of the progressive construction with durative meaning
  - The BE passive, but restricted mainly to durative (v. perfective) constructions

- In Middle English
  - The progressive construction developed its current meaning and drammatically increased in frequency
  - The BE passive expanded to most passive contexts

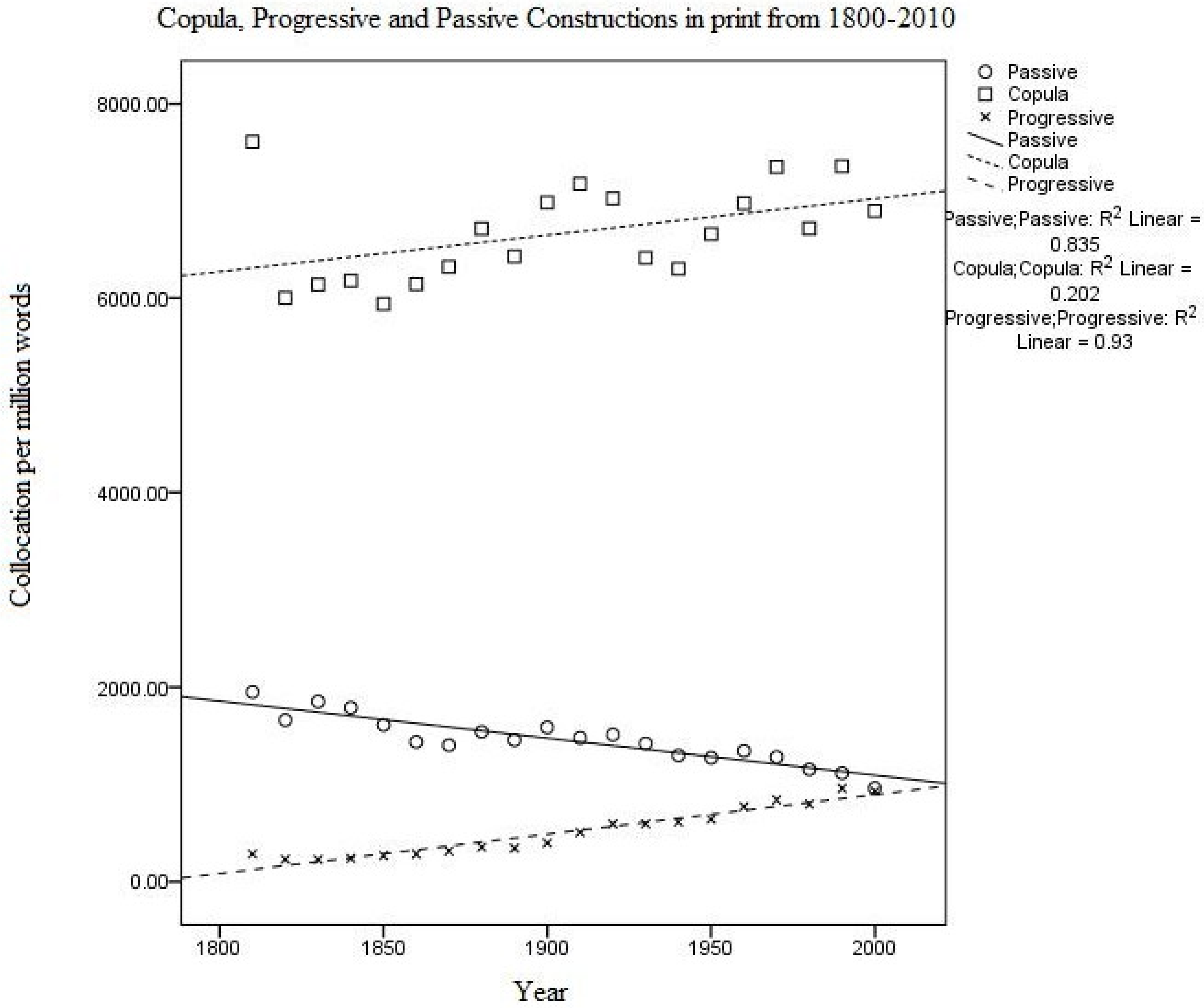
The Constructions in PDE

- In present day English, the progressive construction is increasing in frequency (Leech et al. 2009:121,126)

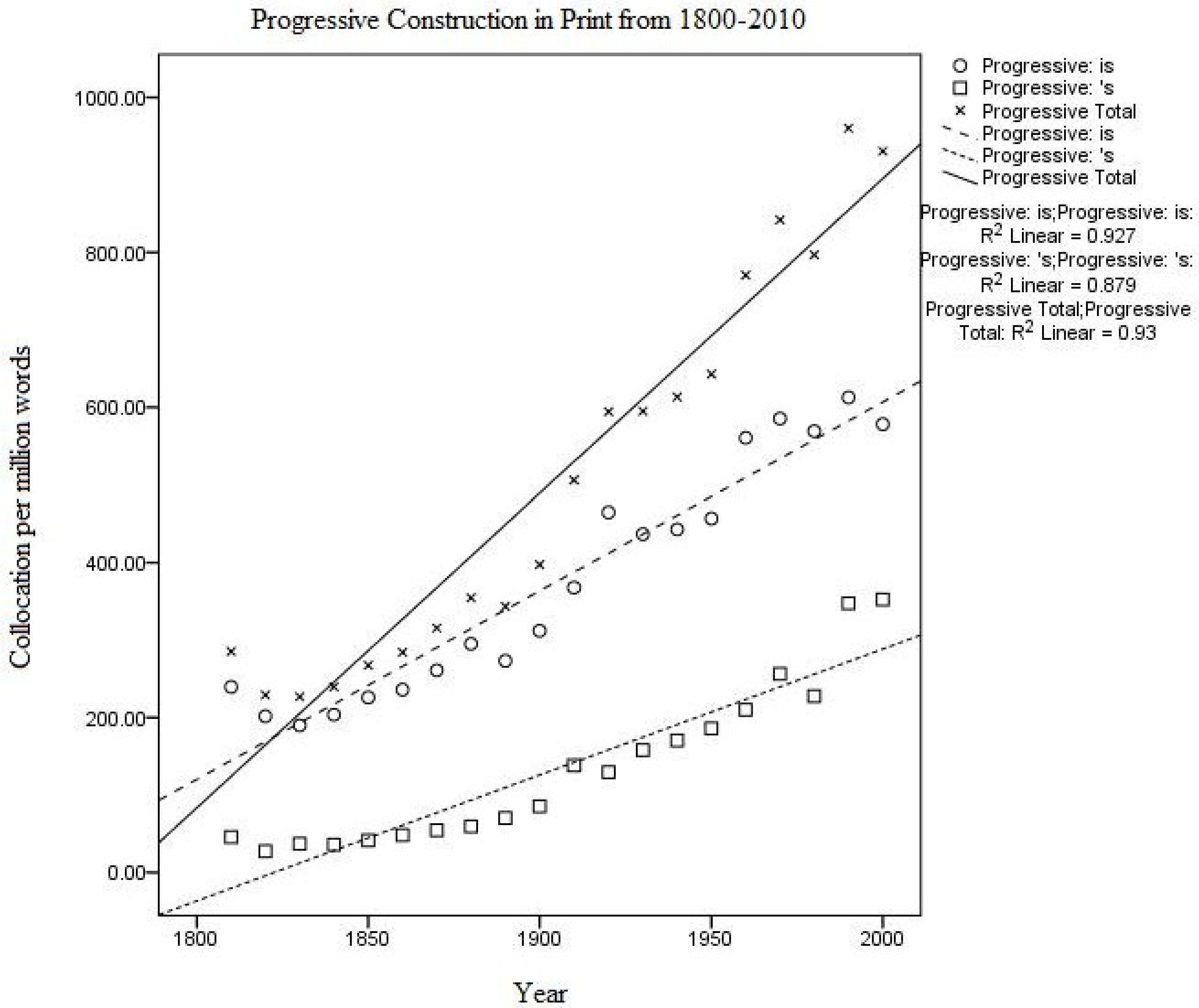
- The BE passive is decreasing in frequency, being replaced by GOT passive (Leech et al. 2009:148)

- This can be seen in COHA (Davies 2010-)

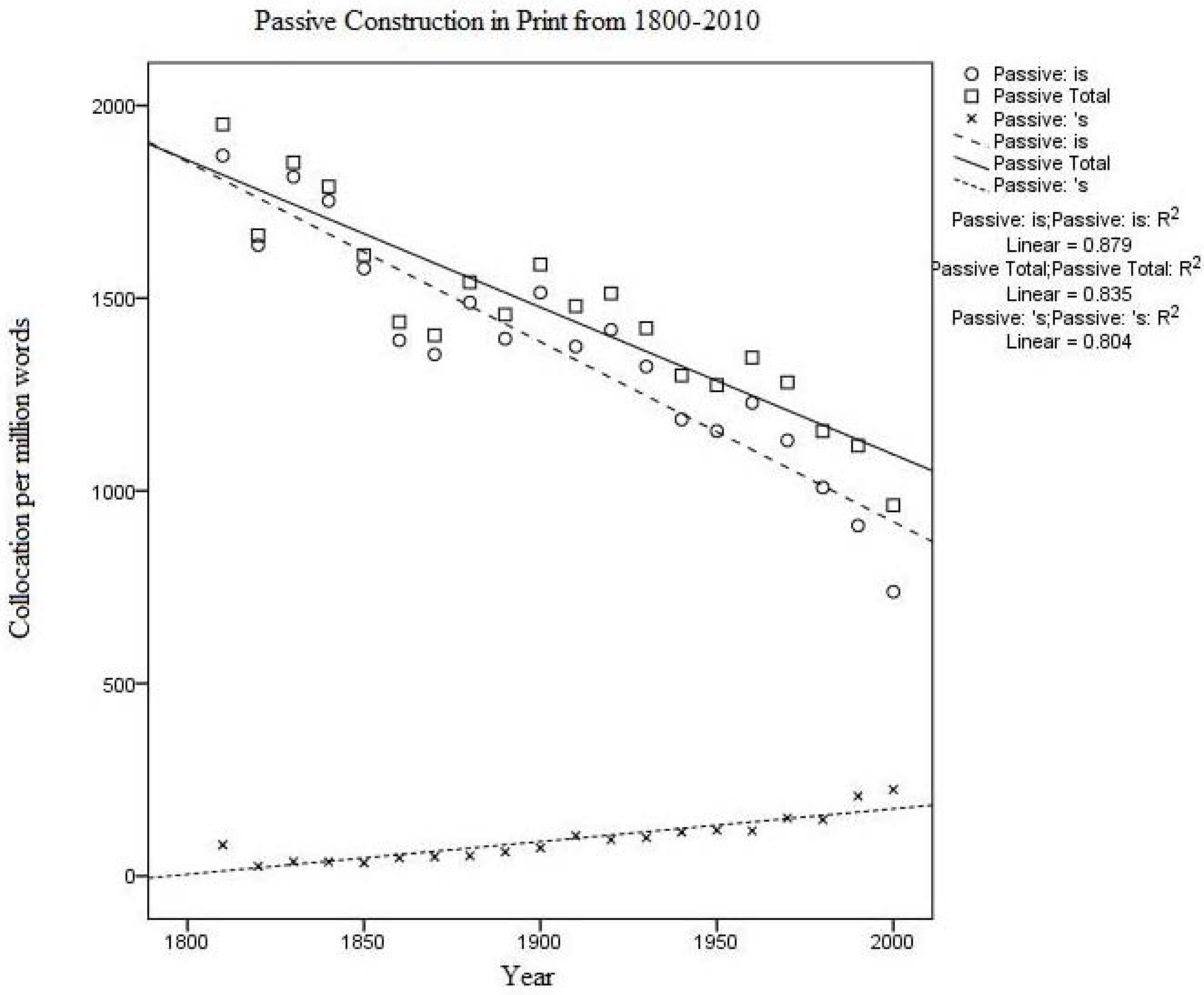
3 is/'s Construction Types in print



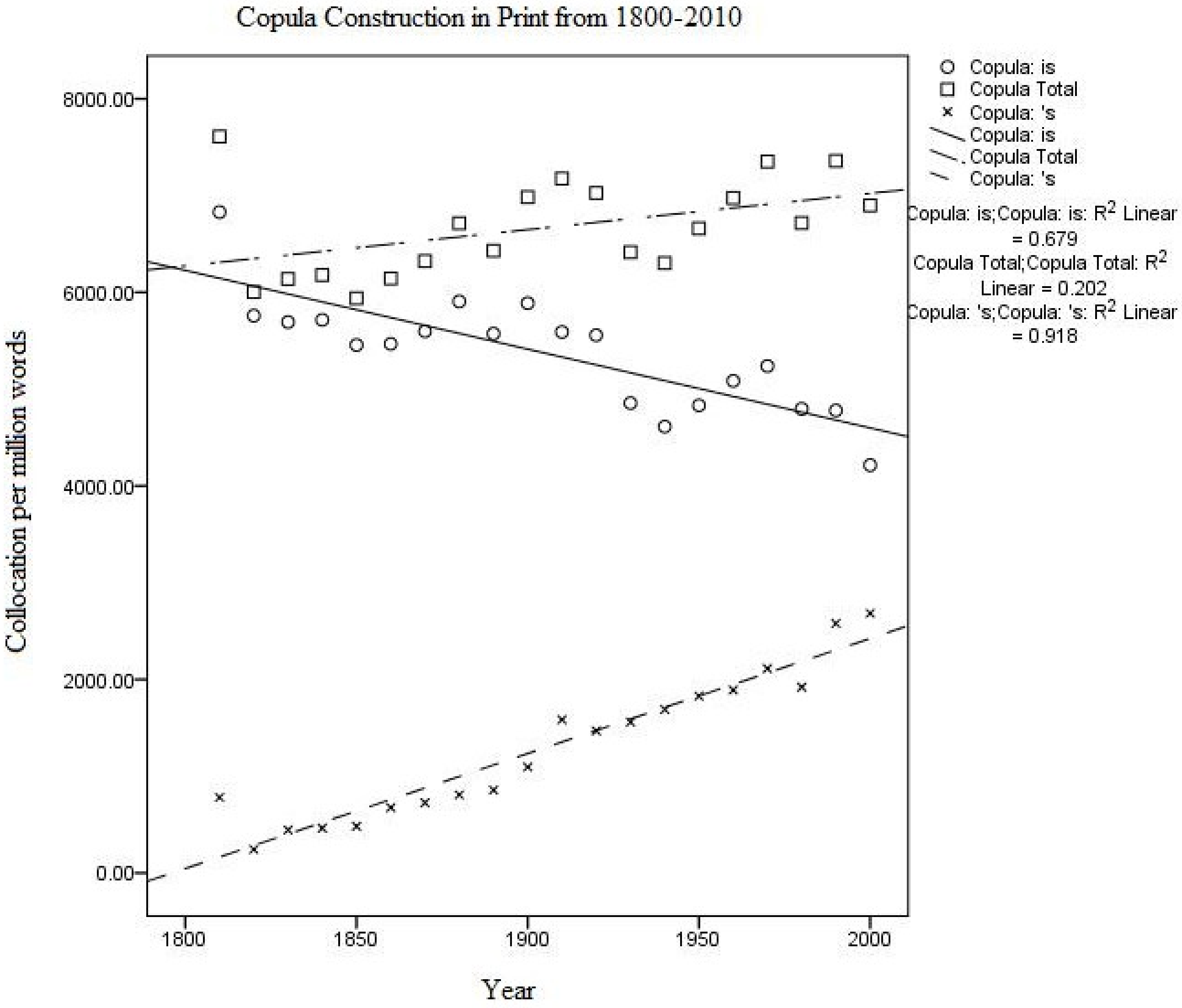
Progressive Construction with is/’s in Print: sharp increase



Passive Construction with is/’s in Print: decrease



Copula Construction with is/’s in print: some increase



Research Question

- Which of the three constructions (copular, progressive, passive) shows the most reduction in spoken (American) English?
    - What factors influence the reduction of the copular, progressive and passive constructions?

The Corpus

- Corpus of Contemporary American English (COCA) (Davies, 2008-)
- Spoken Section has 87,116,763 words (accessed Jan 21, 2011)
- Spoken Section is built from transcripts of live television and radio programs, mostly news programs

Corpus for Model

- A database was created by searching for the targets *is, are, am, ’s, ’re, ’m*
- Approximately 500 entries for each target
- Database reflected overall frequency of construction types in COCA

Construction type	**Copula**	**Progressive**	**Passive**
Reduced: ‘s , ‘m, ‘re	989	544(187)	64
Unreduced: is, am, are	937	371(82)	131

Number of Constructions by Token Types and Construction Types

Excluded tokens

- Tokens were excluded that had:
  - target with a preceding or following disfluency
  - immediate context of target was grammatically incorrect
  - type of construction was not clear
  - ellipsis
  - subject-verb inversion
  - speaker that was unidentifiable
  - for ARE model only: preceding word other than *you, we, they*

Variables – random effects

1. Speaker
2. Show - which program the transcript came from
3. Following phoneme - all vowels were collapsed into one category.
4. Preceding Pronoun - only included in the *is* model, which was only model where there were more than 3 pronouns

Type of Statistical Model

- Logistic mixed-effects model
  - logistic: dependent variable is qualitative not quantitative
  - mixed effects: model has both repeatable/fixed effects and random effects
- Bootstrapping done with a fixed-effects logistic regression model with random effects removed
- Numeric variables were tested for co-linearity
- 4 final models were created: 1 full and 3 individual models for each word form

Testing the Statistical Models

- Factors were added and subtracted to the models to get the best fit
- The simpler model was chosen unless the more complex model accounted for significantly more variance, determined by log-likelihood test
- The Index of Concordance (C) is reported for each model, it measures the concordance between predicted probability and the observed responses
- Significance testing of coefficients through *pvals.fnc* (Baayen 2010).

Results summary

- The progressive construction shows significantly more reduction than the copular and passive constructions
- This is the case even after separating out future constructions, which do not show significantly more reduction than other progressive constructions
- The copular and passive construction do not significantly differ from one another

Results for full model

Construction type	Copula	Progressive	Passive
Reduced: ’re	989	544(187)	64
Unreduced: are	937	371(82)	131

Note. There is a total of 3036 observations in this model, future constructions in parentheses.

- The Passive and Progressive Constructions are significantly different

Results for full model, C = .943

Fixed Factors	MCMC Mean	HPD Lower 95%	HPD Upper 95%	MCMC *p* values
(Intercept)	2.8862	2.7973	2.9752	0.0000
Passive construction (v. Progressive)	**0.0778**	**0.0138**	**0.1377**	**0.0134**
Copula construction (v. Progressive)	0.0281	-0.0061	0.0626	0.1087
Frequency of word string: preceding word and target	**-0.2567**	**-0.2709**	**-0.2430**	**0.0000**
Frequency of word string: target word and following word	**-0.0699**	**-0.0854**	**-0.0589**	**0.0000**
Preceding full BE variant (v. none)	**0.0793**	**0.0448**	**0.1215**	**0.0000**
Preceding reduced BE variant (v. none)	**-0.0670**	**-0.1028**	**-0.0294**	**0.0004**
Preceding unreducable BE variant (v. none)	0.0179	-0.0391	0.0771	0.5397

Random Effects Highlights:

- President Bush, Hillary Clinton, Al Gore and President Obama don’t reduce
- President G. W. Bush, Condoleezza Rice, Bob Dylan and Michelle Obama reduce
- Phonemes most associated with reduction were [l, r, b] and the phonemes most associated with full variants were [ð, v]. These phonemes do not correspond to the most and least frequent following words

Results for IS model

Construction type	Copula	Progressive	Passive
Reduced: ’re	429	81(33)	6
Unreduced: are	411	52 (17)	40

Note. There is a total of 1019 observations in this model, future constructions in parentheses.

- The Progressive Construction is significantly different than the other 2 construction types

Results for IS model, C = .973

Fixed Factors	MCMC Mean	HPD Lower 95%	HPD Upper 95%	MCMC *p* values
(Intercept)	1.9099	1.6591	2.1344	0.0001
Passive construction (v. progressive)	**0.1945**	**0.0856**	**0.3070**	**0.0006**
Copula construction (v. progressive)	**0.0986**	**0.0349**	**0.1605**	**0.0022**
Frequency of word string: preceding word and target	**-0.1101**	**-0.1413**	**-0.0771**	**0.0001**
Frequency of word string: target word and following word	**-0.0208**	**-0.0390**	**-0.0034**	**0.0168**
Preceding full BE variant (v. none)	**0.0756**	**0.0158**	**0.1385**	**0.0178**
Preceding reduced BE (v. none)	-0.0281	-0.0806	0.0229	0.2932
Preceding unreducable BE variant (v. none)	0.0351	-0.0402	0.1137	0.3732
Preceding full NPs (v. non-personal pronouns)	0.2381	-0.1032	0.5855	0.1774
Personal Pronouns (v. non-pers. pronouns)	**-0.3070**	**-0.5739**	**-0.0381**	**0.0242**
Length of preceding NP	**0.0322**	**0.0139**	**0.0513**	**0.0008**

Results for AM model

Construction type	Copula	Progressive	Passive
Reduced: ’re	372	163(57)	19
Unreduced: are	303	125(25)	50

Note. There is a total of 1032 observations in this model, future constructions in parentheses.

- The Progressive Construction is significantly different than the other 2 construction types

Results for AM model, C = .988

Fixed Factors	MCMC Mean	HPD Lower 95%	HPD Upper 95%	MCMC *p* values
(Intercept)	1.7633	1.6618	1.8609	0.0001
Passive construction (v. progressive)	**0.1028**	**0.0134**	**0.1914**	**0.0280**
Copula construction (v. progressive)	**0.1509**	**0.0951**	**0.2084**	**0.0001**
Preceding full BE variant (v. none)	**0.0939**	**0.0309**	**0.1587**	**0.0046**
Preceding reduced BE variant (v. none)	**-0.1060**	**-0.1723**	**-0.0375**	**0.0016**
Preceding unreducable BE variant (v. none)	-0.0196	-0.1179	0.0772	0.6978
Frequency of word string: target word and following word	**-0.0537**	**-0.0782**	**-0.0292**	**0.0001**

Results for ARE model

Construction type	Copula	Progressive	Passive
Reduced: ’re	188	300(97)	39
Unreduced: are	223	194 (40)	41

Note. There is a total of 985 observations in this model, future constructions in parentheses.

- The Copula and Progressive Constructions are significantly different

Results for ARE model, C = .897

Fixed Factors	MCMC Mean	HPD Lower 95%	HPD Upper 95%	MCMC *p* values
(Intercept)	1.6981	1.5621	1.8163	0.0001
Passive construction (v. progressive)	-0.0185	-0.1244	0.0936	0.7408
Copula construction (v. progressive)	**0.0761**	**0.0096**	**0.1445**	**0.0294**
Preceding full BE variant (v. none)	**0.1495**	**0.0747**	**0.2179**	**0.0004**
Preceding reduced BE variant (v. none)	**-0.1017**	**-0.1784**	**-0.0260**	**0.0096**
Preceding unreducable BE variant (v. none)	0.0236	0.0892	0.1331	0.6680
Second person subject (v. third pers. plural)	**-0.2457**	**-0.3145**	**-0.1773**	**0.0001**
First person plural subject (v. third person plural)	-0.0331	0.1044	0.0422	0.3850
Frequency of word string: **target word and following word**	**-0.0405**	**0.0694**	**0.0114**	**0.0062**
Preceding utterance length	**0.0130**	**0.0040**	**0.0216**	**0.0048**

Discussion

- Progressive shows more reduction than other construction types

- The most frequent construction type, copular, never showed the most reduction

Neither frequency or grammaticalization alone have an effect on *is, am,* and *are*

Discussion

- Grammaticalization does put pressure on mid-frequent progressive and future constructions to reduce
- Progressive/Future construction is double marked, making it time intensive for a common pragmatic context -> [aImənə]
- Passive not frequent enough for speakers to experience pressure to reduce, also formal
- Mental representation of passive maybe not fully divorced from representation of copular constructions (partially ambiguous)

Discussion

- Why doesn‘t the copula reduce more often?
- Unlike progressive/passive, the copula is not double-marked
- In focused contexts the copula would be stressed, whereas in progressive/passive the participle would probably be stressed
- From transcripts, it’s impossible to know if this is lexicalized or due to speech conditions
- Data with sound files needed to investigate this further

Discussion: preceding *BE*

- Fowler and Housum (1987) showed that a repeated word is reduced after a first mention
- Here, we get reduced targets associated with reduced previous mentions. Unreduced previous mentions associated with unreduced targets
- Targets probably not second mention
- Could be priming or style matching
- Speaker as a random variable should have factored out some of the noise from certain people just being more likely to use reduced or unreduced variants.
- Also preceding *BE*s could come from another interlocutor (cf. Show as random variable)

Discussion: collocate frequency

- Word string frequency is discussed by Bybee and Scheibman (1999) as a predictor of reduction
- This variable preformed better than two other types of frequency: conditional probability (Bell et al. 2009), log frequency of collocate
- Conditional probability was also significant, but word string frequency preformed better in log-likelihood tests
- The preceding context had a stronger coefficent than the following context

Discussion: Pronouns

- Personal pronouns far more likely to occur with reduced variants
- From random effect we know that the individual pronouns most associated with *'s* were *here* and *what* (despite not being personal pronouns)
- Pronouns most associated with *is* were *this* and *which* (these end in sibilants, but preceding sibilant was not a significant factor in the model)

Future research

- Use spoken corpus to find *’re* with other NPs than *you, we, they*
- Using finer measures of reduction: duration measurements from a spoken corpus, laboratory experiment
- Comparing reduction in a contraction-licensed language (English) and a non-contraction-licensed language (German)
- Comparing reduction in verb-aux pairs where verb does not reduce (*have~’ve, has~’s*)

References

Bell, A., J. Brenier, M. Gregory, C. Girand and D. Jurafsky. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60 (1), 92-111.

Bybee, J. L. (2007). *Frequency of use and the organization of language*. Oxford; New York: Oxford University Press.

Bybee, J. & Pagliuca, W. (1985). Cross-linguistic comparison and the development of grammatical meaning. In Fisiak (ed.), 60-83.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H. (2010). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". R package version 1.0. http://CRAN.R-project.org/package=languageR

Burrow, J. A. and T. Turville-Petre (1996). *A Book of Middle English.* Oxford: Blackwell Publishers Ltd.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 400+ million words, 1990-present. Available online at http://www.americancorpus.org.

Davies, Mark. (2010-) The Corpus of Historical American English (COHA): 400+ million words, 1810-2009. Available online at http://corpus.byu.edu/coha.

Fowler, C. A., Housum, J. (1987). 'Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction.' *Journal of Memory and Language*, 26, 489-504.

Gabelentz, G. (1891). *Die sprachwissenschaft, ihre aufgaben, methoden und bisherigen ergebnisse*. Leipzig,: T. O. Weigel nachfolger.

Gahl, S. (2008). 'Time' and 'thyme' are not homophones: The effects of lemma frequency on word durations in spontaneous speech. Language, 84, 474-96.

Givón, T. (1985). Iconicity, isomorphism, and non-arbitrary coding in syntax. In Haiman, J. (ed.), 187-219.

References cont.

Heine, B., (1993). *Auxiliaries*. Oxford: Oxford University Press.

Heine, B., Claudi, U., & Hünnemeyer, F. (1991). *Grammaticalization: A conceptual framework*. Chicago: The University of Chicago Press.

Hopper, P. and Traugott, E. (1993). *Grammaticalization*. Cambridge: Cambridge University Press.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: The John Benjamins Publishing Company, 229-254.

Leech, G., M. Hundt, C. Mair and N. Smith (2009). *Change in Contemporary English*. Cambridge: Cambridge University Press.

Lehmann, C. (1995). *Thoughts on Grammaticalization*. München, Newcastle: Lincom Europa.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In Bybee, J. and P. Hopper (eds) *Frequency effects and the emergence of linguistic structure.* Amsterdam: The John Benjamins Publishing Company, 137-157.

Pierrehumbert, J. (2002). Word-specific phonetics. In Gussenhoven, C. and N. Warner (eds), *Laboratory phonology VII (phonology and phonetics)*. Berlin: Mouton de Gruyter, 101-140.

Quirk, R. and C. L. Wrenn (1957). *An Old English Grammar*. London: Methuen & Co Ltd.

R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

van Bergem, D. (1995). *Acoustic and lexical vowel reduction*. Amsterdam: IFOTT.

Visser, F. Th. (1963-1973). *An Historical Syntax of the English Language*. 3 vols. Leiden: E. J. Brill.

Copula Construction in OE

- The copula construction was present in Old English:

Ic ***beo*** *mid eow ealle dagas*

‘I **am** with you always’

Gospel Matthew 28:20 cited by Visser (1963:160)

- Has not changed greatly since then: same syntactic position, same complements – adjectival, nominal, preopositional

Progressive in OE

- One option for expressing a durative meaning was the forerunner of the progressive – BE + present participle with <ende>

ic mē gebidde to ðǣm Gode þe ***bīō eardigende*** *on heofonum*

'I pray (at this moment) to the God who **is dwelling** (not only at this moment) in the heavens' (Quirk and Wrenn 1957:80).

Progressive in ME

- Became more frequent, <ende> became <ing/ung>, perhaps due to analogy with gerunds in locative constructions, i.e. ‘he is on huntung’, progressive meaning

Heo...iuunden Þene king Þær he ***wes an slæting***

‘and they found the king where he **was hunting**’

Layamon’s Brut cited by Visser (1966:1095)

Passive in OE

- One option for expressing a passive was BE + past participle, used mostly with durative constructions, BECOME passive used with perfective constructions, but great deal of variation (Quirk and Wrenn 1957:80-81).

Ne ***bið*** *ðǣr nǣnig ealo* ***gebrowen***

'No ale **is** (ever) **brewed** there‘

(Quirk and Wrenn 1957:80)

Passive in ME

- Most passives in ME were now expressed with BE auxiliary

*he...**wæs** wæl **underfangen** fram Þe pape Eugenie*

'He **was** well **received** by Pope Eugenius' (Burrow and Turville-Petre 1996:52)

Variables

1. Construction Type – Copula, Progressive or Passive
2. Occurrence of Preceding BE in 9 preceding words – Full BE (*is, am, are*), Reduced BE (*’s, ’m, ’re*), Unreducable BE (*be, being, been, was, were*), None
3. Log frequency of word string: target word and following word

Variables

4. Log frequency of word string: preceding word and target
5. NP Type – personal pronoun, non-personal pronoun, non-pronominal
6. Length (in words) of preceding NP
7. Length (in words) of preceding utterance
8. Subject – third person plural, first person plural or second person