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
  Grammaticalization 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 it to be so
  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?


- 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

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

*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 dramatically 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

Copula, Progressive and Passive Constructions in print from 1800-2010

- Passive
- Copula
- Progressive
- Passive
- Copula
- Progressive

Passive Passive: R^2 Linear = 0.835
Copula Copula: R^2 Linear = 0.202
Progressive Progressive: R^2 Linear = 0.93
Progressive Construction with is/’s in Print: sharp increase
Passive Construction with is/’s in Print: decrease

Passive Construction in Print from 1800-2010

- Passive: is
- Passive Total
- Passive: ’s
- Passive Total
- Passive: ’s
- Passive Total

Passive: is
1. Linear = 0.879
   R^2 = 0.771

Passive Total
1. Linear = 0.635
   R^2 = 0.400

Passive: ’s
1. Linear = 0.804
   R^2 = 0.646

Collocation per million words

Year
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

<table>
<thead>
<tr>
<th>Construction type</th>
<th>Copula</th>
<th>Progressive</th>
<th>Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced: ‘s, ‘m, ‘re</td>
<td>989</td>
<td>544(187)</td>
<td>64</td>
</tr>
<tr>
<td>Unreduced: is, am, are</td>
<td>937</td>
<td>371(82)</td>
<td>131</td>
</tr>
</tbody>
</table>

*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

<table>
<thead>
<tr>
<th>Construction type</th>
<th>Copula</th>
<th>Progressive</th>
<th>Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced: 're</td>
<td>989</td>
<td>544(187)</td>
<td>64</td>
</tr>
<tr>
<td>Unreduced: are</td>
<td>937</td>
<td>371(82)</td>
<td>131</td>
</tr>
</tbody>
</table>

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

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

**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

<table>
<thead>
<tr>
<th>Construction type</th>
<th>Copula</th>
<th>Progressive</th>
<th>Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced: 're</td>
<td>429</td>
<td>81 (33)</td>
<td>6</td>
</tr>
<tr>
<td>Unreduced: are</td>
<td>411</td>
<td>52 (17)</td>
<td>40</td>
</tr>
</tbody>
</table>

*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

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

<table>
<thead>
<tr>
<th>Construction type</th>
<th>Copula</th>
<th>Progressive</th>
<th>Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced: 're</td>
<td>372</td>
<td>163(57)</td>
<td>19</td>
</tr>
<tr>
<td>Unreduced: are</td>
<td>303</td>
<td>125(25)</td>
<td>50</td>
</tr>
</tbody>
</table>

*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

<table>
<thead>
<tr>
<th>Fixed Factors</th>
<th>MCMC Mean</th>
<th>HPD Lower 95%</th>
<th>HPD Upper 95%</th>
<th>MCMC p values</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.7633</td>
<td>1.6618</td>
<td>1.8609</td>
<td>0.0001</td>
</tr>
<tr>
<td>Passive construction (v. progressive)</td>
<td>0.1028</td>
<td>0.0134</td>
<td>0.1914</td>
<td>0.0280</td>
</tr>
<tr>
<td>Copula construction (v. progressive)</td>
<td>0.1509</td>
<td>0.0951</td>
<td>0.2084</td>
<td>0.0001</td>
</tr>
<tr>
<td>Preceding full BE variant (v. none)</td>
<td>0.0939</td>
<td>0.0309</td>
<td>0.1587</td>
<td>0.0046</td>
</tr>
<tr>
<td>Preceding reduced BE variant (v. none)</td>
<td>-0.1060</td>
<td>-0.1723</td>
<td>-0.0375</td>
<td>0.0016</td>
</tr>
<tr>
<td>Preceding unreducible BE variant (v. none)</td>
<td>-0.0196</td>
<td>-0.1179</td>
<td>0.0772</td>
<td>0.6978</td>
</tr>
<tr>
<td>Frequency of word string: target word and following word</td>
<td>-0.0537</td>
<td>-0.0782</td>
<td>-0.0292</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
Results for ARE model

<table>
<thead>
<tr>
<th>Construction type</th>
<th>Copula</th>
<th>Progressive</th>
<th>Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced: ’re</td>
<td>188</td>
<td>300 (97)</td>
<td>39</td>
</tr>
<tr>
<td>Unreduced: are</td>
<td>223</td>
<td>194 (40)</td>
<td>41</td>
</tr>
</tbody>
</table>

*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

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


Copula Construction in OE

- The copula construction was present in Old English:

  \( Ic \texttt{beo} \texttt{mid} \texttt{eow} \texttt{ealle} \texttt{dagas} \)

  ‘I \texttt{am} with you always’


- Has not changed greatly since then: same syntactic position, same complements – adjectival, nominal, prepositional
Progressive in OE

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

\textit{ic mē gebidde to Ȝām Gode þe bīō eardigende on heofonum}

'I pray (at this moment) to the God who \textbf{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 ðær nænig 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...was well underfangen from the 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), Unreducible 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