Variation and change in the productivity of BE *going to* V

in the *Corpus of Historical American English*, 1810–2009

Tanja Säily, Florent Perek & Jukka Suomela

Grammaticalization of BE going to V

- 1. I'm going to the market to buy bananas
 'motion with intention'
- 2. *I'm going to read your work tomorrow* 'motionless intention'; EModE
- 3. There's going to be some serious trouble here 'prediction'; LModE-PDE
 - a. You're going to **feel** very foolish (mental verb; COHA, 1932)
 - b. It's going to rain (inanimate subject, it; COHA, 1811)
 - c. Father Paul was going to **be cheated** of his share (passive voice; COHA, 1946)

(Budts & Petré 2016; Wu et al. 2016)

Research questions

- 1. How is the grammaticalization reflected in the **productivity** of the construction in LModE-PDE?
 - Internal factors: semantics of the verb (including mental verbs), inanimate subject (*it*), passive voice
- 2. Did the social factor of **gender** play a role in the process?

Material

- Corpus of Historical American English (COHA)
 400 Mw, 1810-2009
- Fiction section: c. 50% of the data
 - Gender metadata for authors developed by Öhman et al. (2019)
 - Promising material for sociolinguistic investigation: a more speech-like genre (dialogue)
 - Types of fiction (e.g. short stories, drama, movie scripts) unevenly distributed over time (Säily & Vartiainen forthcoming)
 → restriction to novels only, c. 150 Mw
- List of mental verbs from Halliday & Matthiessen (2014: 256–257)

Analysis 1: type frequencies

Methods

- We study the productivity of BE going to V by studying type frequencies
 - I.e. how many different verbs follow BE going to in different time periods
- Key challenges:
 - Different amounts of text from different time periods, different amounts of text from men and women: how to compare type frequencies?
 - If we observe trends, are they **statistically significant**?



Choose **random subcorpora** with the same number of words from each time period

Visualizing trends



Assessing statistical significance









6

0

ō σ

ŝ

Significance of differences in time



More frequent use or more diverse use?

2009

1

0

-

σ

σ

ŝ σ

6 ō

 σ

0 \circ 0

94 LO 9 67

9

6 σ

30

6





Time period

Time period

Time period

Type frequency

- Slight decrease in types over time, especially for men
- Type counts for men and women converge over time
- But what *kind* of types are they?
 - \circ $\,$ Is the construction undergoing semantic specialization?
 - \circ $\,$ Do men and women use it in different semantic areas?
- We examine these questions using *distributional semantics*

Analysis 2: distributional semantics

Distributional semantics

"You shall know a word by the company it keeps" Firth (1957: 11)

- Aim = capturing word meaning through lexical collocates in large text corpora
- Semantically similar words are expected to have the same collocates

• e.g. drink and sip > wine, water, coffee, cup, bottle, etc.

• Semantic similarity is approximated by similarity in distribution

Distributional semantic model

- DSM built with word2vec (SkipGram, cf. Mikolov et al. 2013), using gensim
- Trained on the whole COHA, context window +/- 2 words
- Each word is assigned a "vector", i.e. array of values
- This quantification of meaning allows us to (*inter alia*):
 - Visualise the semantic distance between a set of words by plotting them in two dimensions (using e.g. t-SNE) (Perek 2016, 2018)
 - Measure and compare the semantic spread of constructions (Hilpert & Perek 2022)

Distributional semantic plots, whole corpus

Using random samples (matched for frequency, N = 534)

1820-1859 1860-1899 1900-1939 1940-1979 persuade convince 9966P advisera red Mineroforbid propose independence prevent force Aut and argue sentrable sonance seek dd RXBBd suggesteed stop announdestant and beg finish happen happen effect happen offer attach bservebefall learn supply loarn relatemmit marry marry marr recortifice marry saine Comp 9 (9 mg/s appoint resign te segdanalenv divorce state pray obe^dicelike search explote adjourn t...testnænsbratsforsW8^h (Alaa atax Sileyn visit goverroduce aller 00/ prongynge obtair arv scream c /6/2/am enformi establish support transact P@Xn been P®Vin built rick Billin buy incue simply incue crosk play hola at (BV) - defend build Men hire handle inera a ch fight < 91/0 <38/e sairpas centrised of separate recoasete optista sideback wak secede s pisa é spland colors s feldp 信息 sten mount satiafitaicape catch overtake ekidena dis elses tunnegen: vacate nurry chase disquise scolo accept rract. rus. reform faint remake STREZ P - buret lip roll wreiri ruin destroy increase _ laygh tor change dia in the second s Karve perjure drown imprison die arres t slan trim oldesider hurt affect hurt rokill wake rokill musider tack rummigae disturb dean% _{desert} DI S JERRI ASISI lick pick drink unde wiperash inhabit swamp wipe fill 001/0 pack crumble asfika mend repai r899t roast 1820-1859 1860-1899 1900-1939 1940-1979 inquirar6899* order add insist agree allow prevent allow choose s trips and the Binvite succeed to announcestask a <mark>ficish</mark>olish happer happen oroverespond Women marn marrie marry Haiss applyannes rite send roite land alari denant alant resian - Kieto waktetp like oiexplore doint question /e postpone meet, technikten forgiver? ataxia presideblesstrust restudy quotesttend Side Week aovern soleam av conduct gain 5400 pay tand act teachdild^{or os} summin buy serve practis e 6.060 serve **bux** teachuild win buy depart puttorit receive last redeanve :9,81/B flymp2885 quarrel onst flirt ounish hallenge weitige cable os e will be guide disinhari SDIEBE splead skiptar df9986dgeig ria core evade riðði^{lþalkatis fy} catch sling escape indulas, mount unite accept accept accept storm faint faint terfaint stick blaze revive reconnoitrarophr reconnoitre rattle nuin spoil ruin blugmit settle laugh fall destroy -96946 stresknburn change change the sine disrebange diarve dia nainhf buRia exhaust affect hurt kies scalp^{by}fangle_{kill} deprimerder < rewarder waklede kill e lit bother pick & gellow pid and desert elativour 6.9 entestariorise impress aningsy stuff dump pack **algyt**k

shine broil

cook

Using random samples #2 (N = 534)

Using random samples #3 (N = 534)

1820-1859 1860-1899 1900-1939 1940-1979 persuade all the pareforbid advise advise reply remonstratepose reply propose prevent insist seek refasee fail rofuse seek obligg Stop Snnounce manage end uggesteed and suggest happen command confees consistemate beai happen offect hanner leabefall offer commit howlerstand marry marry marry sacrifice **ман** у present end send introduce m dabak write send elect retire send 4ver.ee resign issue _{quiz}adm**iniele**th likelish ohoteereehteril heedike namte or for give guaraduce quate goverroduce procure trust if#aay SUBALIES staa scream s 🕬 am conduct perform maintain appea appear sing blav invertablis! dispose e coto croaking leve. croaking P∜¥in ₩₩ lodge bunkefend tendagt engage w^b riskeiter sell teachre engagech Men handla teach inflict save fight lifes t fight fatch 100000 parallel rescue coonsisters free release r ußarr nunisl sauntevenaneuve stepp rai solate llenge white pilot_{draw} spland s close e asolged ii) catdicape foreclose accept raise 148 saties sneak chase bose scold Helf8rm storm ush, hide defraud rej**al**arae stoppinp faint fairt iumo stickme explodesRi run lear bounce linflop rebuild papil assist divide wnairk reduce change destro laugh emove change hoa bugle imprison hang toss waar diarve rotarve choke taulper hurt freeze mugger affect tap tear trim affect olonder grip fridfiteHrb trouble astonism trouble rub pick detentk rummeaert plucken pleaseffend squeeze eat OVERFIND occupy ease wip@ash deposit inhabit ease stuff fill pack ANEd repai bake bake 1820-1859 1860-1899 1900-1939 1940-1979 inform inquire propose allow allow add demagd ek h chotetuse seek stop anngungest afford end need mis carry criticieto happene happen happen 6/e show forget All shokelieve ainlearr relate JWAVE marrie sent marry introduce naerry Women The states introduce wette sen dalkok alact divorce - MAR dial de la likeniov weep eniov likeniov tiken consult attend Istaway! texe n attend govern 193 198 scielan protect GRAW1 perform collect treat pay sing Pay and the second Payin blogs specialize gegeechuild fulfill win buy 560 teachuild play teach receive ncentrate bear PURCHR reward last thunde confro fight Shir Cass board SAVA Jang^{leee} free SalPAYE quarrel - onst nieh excel shiea cable cable wandeidee rambleite weide spend difue interfere Faise taging hauraise satisfy s atis fy catch catch sling supersede mutterool accept accept faint stowne faint stick nobneden blaze form reconnoitre rattle ruin slip ruin roll rain strikent settle laugh DISH cultivate change throwsnap stifteet contenetange IN CIVIL throw rap* pave die deatrate builte diarve ween pediah arrest harm smash kiss hurt drill burn hud n kişs hurt tais deprimerder munder weave kill trouble bother nick eat elativour eat eat anhard swamp witelean dout nack герани bake shine broil

Qualitative type-based analysis

- Type distribution highly variable from sample to sample
- Problem = we cannot average over individual types!
 - \circ $\;$ But we can average over type counts $\;$
 - \circ $\,$ We just need to add a semantic dimension to type counts $\,$
- Idea: types are sorted into discrete semantic categories
 - \circ We can average over type counts in each category across samples
 - $\circ~$ This gives us a representation of the average "semantic spread" of the construction

Qualitative type-based analysis

- We collect all types in the random samples (1419 types)
- We extract pairwise semantic similarity scores between these types from the DSM
- We use these scores to automatically group types into semantic categories using cluster analysis (PAM)

Qualitative type-based analysis

- In each cluster, we calculate the average number of types attested in each period across the 1000
- Similarity between type distributions can be measured using Pearson's correlation coefficient (ρ)
 - Between different periods
 - \circ $\,$ Between genders in the same period $\,$

	1820-1859	1860-1899	1900-1939	1940-1979
Cluster 1	11.467	10.841	10.885	10.626
Cluster 2	6.434	5.892	6.958	7.35
Cluster 3	19.707	19.289	18.35	16.016
•••	····	····	•••	•••
	$\rho = 0$			

Variation in semantic spread (30 clusters)

Variation in semantic spread (200 clusters)

Proportion of types per cluster for MEN (30 clusters)

Proportion of types per cluster for WOMEN (30 clusters)

Time

Time

Cluster 28: mental verbs (cognitive type)

learn accept feel believe forget explain satisfy remember discover suit regret understand ignore mind realize fear penetrate unravel suspect guess recognize appreciate heed interpret describe wonder respect judge notice acquaint suppose

(+ stop mention solve belong materialize exist excuse ...)

Corpus token frequency for the cluster 28 types

Cluster 3: verbs of motion

leave run drive ride walk follow move dance sail enter cross travel step march head hurry hike wander wade trot sneak swim stumble meander shuffle stride stroll approach descend skate parade trudge ramble circle tread amble paddle stalk saunter sprint

(+ stay live play carry stand pass sit wait lead watch listen row usher face ...)

Discussion

Summary of results

- Overall productivity/type diversity of BE going to V doesn't increase in C19-20 AmE, even a slight decrease

 Men's usage more productive, convergence over time
- Internal factors do indicate increasing productivity
 - Proportion of types with mental verbs, *it* subjects, passive voice
 - \circ $\,$ Women lead increase in mental verbs $\,$
- Type-based semantic analysis identifies areas of growth
 - E.g. mental verbs, motion verbs
 - \circ $\,$ Points to an increase in grammaticalization $\,$
 - \circ Gender differences as well, with women leading the way

Conclusions

- At this stage of grammaticalization, overall type diversity stagnates but **internal factors** linked to grammaticalization indicate increasing productivity
 - \circ Important to take into account
- Consistent **gender** differences different leaders of change and/or different genres?
 - \circ $\,$ Gender cannot be ignored as a possible factor $\,$
 - Mental verbs could be linked to women's involved writing style (Biber & Burges 2000)
- Future work: analyse hapax legomena / new types

References

- Biber, Douglas & Jená Burges. 2000. Historical change in the language use of women and men: Gender differences in dramatic dialogue. Journal of English Linguistics 28(1): 21–37.
- Budts, Sara & Peter Petré. 2016. Reading the intentions of *be going to*: On the subjectification of future markers. *Folia* Linguistica Historica 37: 1–32.
- Halliday, M.A.K. & Christian M.I.M. Matthiessen (2014). Halliday's introduction to functional grammar, 4th edition. Routledge.
- Hilpert, Martin & Florent Perek. 2022. You don't get to see that every day: On the development of permissive get. Constructions and Frames 14(1): 14-41.
- Öhman, Emily, Tanja Säily & Mikko Laitinen. 2019. Towards the inevitable demise of *everybody*? A multifactorial analysis of *-one/-body/-man* variation in indefinite pronouns in historical American English. 40th Annual Conference of the International Computer Archive of Modern and Medieval English (ICAME 40), Neuchâtel, Switzerland, June 2019. https://tanjasaily.fi/talks/icame40 ohman et al 2019.pdf
- Perek, Florent. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics* 54(1): 149–188.
- Perek, Florent. 2018. Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. Corpus Linguistics and Linguistic Theory 14(1): 65–97.
- Rodríguez-Puente, Paula, Tanja Säily & Jukka Suomela. 2022. New methods for analysing diachronic suffix competition across registers: How -*ity* gained ground on -*ness* in Early Modern English. *International Journal of Corpus Linguistics* 27(4): 506-528.
- Säily, Tanja, Martin Hilpert & Jukka Suomela. Forthcoming. New approaches to investigating change in derivational productivity: Gender and internal factors in the development of -ity and -ness, 1600-1800. In Patricia Ronan, Theresa Neumaier, Lisa Westermayer, Andreas Weilinghoff & Sarah Buschfeld (eds.), Crossing boundaries through corpora: Innovative approaches to corpus linguistics. Benjamins.
- Säily, Tanja & Turo Vartiainen. Forthcoming. Historical linguistics. In Michaela Mahlberg & Gavin Brookes (eds.), Bloomsbury handbook of corpus linguistics. Bloomsbury.
- Wu, Junhui, Qingshun He & Guangwu Feng. 2016. Rethinking the grammaticalization of future be going to: A corpus-based approach. Journal of Quantitative Linguistics 23(4): 317–341.

What about gonna?

-2009 1820-1859 σ Н H H Н F H H -Ч 1830-1850-2 1870-1810-1920-1970-H

σ 5

-2 m 4 95

6 6 19 0

--H 969

0

σ

~ õ

5

5

-Ē -H

6

σ

σ

-2009

970

σ

5

5 σ

-

0 \circ 96

თ

Ь

0

859 1869 1879

Т

1820-

1 1 1 1 1

830-

840 1850 860 870 880 .890 900 10 920 30 940

1889 1899 606

849

--

0

-

ŝ

1970-2009

849 869 1879 1889 1909 949 666 1820-1859 σ σ σ 1970-2009 89 92 63 95 96 67 Г 98 5 Ч L Ч Г H H Г Ч H H -Ч 1830-7 1850-2 1870-3 1 1 1 1 1 1 1 1 1880-1900-1810 1840 1860 1890 1910 1920-30 1940 1950 1960 6 H

849 1869 1879 666 1820-1859 1889 1870-1909 σ σ σ 1970-2009 89 92 63 94 95 96 67 Г 86 5 Ч Н Г H H Г F H H -Ч 1830-7 1850-2 1900-7 1 1880-1 1 1 1 1 1810 1840 1860 1890 1910 1920-930-1940 1950 1960 H