Usage-based testing of linguistic hypotheses: The case of Dutch causatives

Dirk Geeraerts

University of Leuven

RU Quantitative Lexicology and Variational Linguistics

Purpose

present an illustration of a lectally enriched multivariate usage-based grammar, i.e. a non-modular corpus-based description of linguistic behaviour which specifically incorporates pragmatic factors of a variationist and discursive nature

Assumptions

1 language behaviour takes the form of choices between alternative construals;

therefore, linguistic description in general takes the form of an onomasiological analysis of those choices

Assumptions

2 choices of this kind are determined by the interaction of diverse factors of a structural, referential, discursive, lectal nature;

therefore, it would be counterproductive to separate ‘pragmatic’ factors from others (conceptually speaking, it is a possibility, but descriptively speaking, all the factors work simultaneously and interactively)

Assumptions

3 studying such the multifactorial nature of language behaviour requires an appropriate methodology, i.e. a multivariate statistical analysis of large samples of actual language use

therefore, linguistic description involves empirical testing of hypotheses concerning language use

Background

Speelman, Dirk and Dirk Geeraerts. 2009. Causes for causatives: the case of Dutch 'doen' and 'laten'. In Ted Sanders and Eve Sweetser (eds.), Causal Categories in Discourse and Cognition. Berlin: Mouton de Gruyter.

[José Tummers, Kris Heylen & Dirk Geeraerts. 2005. “Usage-based approaches in Cognitive Linguistics. A technical state of the art”. Corpus Linguistics and Linguistic Theory 1(2).]

Background

other applications include

- word order phenomena (De Sutter)
- particles (Grondelaers)
- flectional variation (Tummers)
- loans and borrowings (Zenner)
- color terms (Anishchanka)
- metonymic patterns (Zhang)
- diachronic change (Gevaert)

What ?

1° is there a significant distinction between Belgian Dutch and Netherlandic Dutch re the distribution of doen and laten ?

terminological shorthand:

- Belgian Dutch : Flemish
- Netherlandic Dutch: Dutch
- variation involving dialects, regiolects, idiolects, sociolects, register, national varieties etc.: lectal variation

What ?

2° is the distribution of laten determined by indirect causation ?

Kemmer & Verhagen 1994; Stukker 2006:

direct causation: doen

indirect causation: laten

What ?

2° is the distribution of laten determined by indirect causation ?

Kemmer & Verhagen 1994; Stukker 2006:

direct causation: doen

indirect causation: laten

"The causer produces the effected event directly; there is no intervening energy source 'downstream'"
(Stukker 2006: 50)

What ?

2° is the distribution of late

Kemmer & Verhagen 1994,

direct causation: doen

indirect causation: laten

"Besides the causer, the causee is the most immediate source of energy in the effected event. The causee has some degree of 'autonomy' in the causal process"
(Stukker 2006: 50)

How ?

what we need:

- a representative corpus of language data
- a set of potentially relevant factors coded in the corpus
- a statistical technique analysing the relevance of the factors

How ?

what we need:

- a representative corpus of language data
- a set of potentially relevant factors coded in the corpus
- a statistical technique analysing the relevance of the factors

CGN – Corpus of spoken Dutch, release 1.0

900 hrs, tagged

2/3 Dutch, 1/3 Flemish

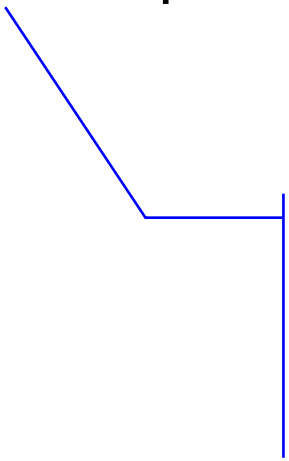
register and text type variation (see below)

automatic data selection with manual correction

How ?

what we need:

- a representative corpus of language data
- a set of potentially relevant factors coded in the corpus
- a statistical technique analysing the relevance of the factors



stepwise logistic regression:

what is the impact of a multitude of possibly relevant factors

on the variation observed in the (categorical) data?

The external factors

variation wired in into the CGN:

- speaker characteristics: sex, age, educational level
- regional variation: Flemish vs. Dutch
- register variation: 15 'components', divided along three dimensions

dialogues and multilogues vs. monologues

private speech vs. public speech

spontaneous vs. prepared speech

The external factors

spontaneous:

A face-to-face conversations

B interviews (with teachers)

C spontaneous telephone conversations (recorded via switchboard)

D spontaneous telephone conversations (recorded locally)

E simulated business negotations

F broadcast interviews/discussions/debates

H classrooms lessons

I live (sports) commentaries

prepared:

G (non-broadcast) political discussions and debates

J broadcast newsreports and reportages

K broadcast news

L broadcast commentaries and reviews

M ceremonious speeches and sermons

N lectures and seminars

O written texts read aloud

The internal factors

- syntactic construction type
- coreferentiality between matrix subject and infinitival subject/object
- animacy of matrix subject
- lexical collocational strength
- conceptual collocational strength

PS why these ? for theoretical reasons, and on the basis of an exploratory scanning of the data;
it is customary in regression analysis to start with a broad set of parameters, and then to reduce it,
automatically as a result of the regression, and manually by considering different ways of coding

The internal factors

- syntactic construction type
- coreferentiality between mat subject and infinitival subject/object
- animacy of matrix subject
- lexical collocational strength
- conceptual collocational stren

PS why these ? for theoretical reason
it is customary in regression analysis t
automatically as a result of the regressio

if laten expresses indirect causation, you don't expect laten in intransitive constructions, where there is no intermediate entity

The internal factors

- syntactic construction type
- coreferentiality between matrix subject and infinitival subject/object
- animacy of m[illegible]ix subject
- lexical collocational [illegible]th
- conceptual collocational s[illegible]

PS why these ? for theoretical reasons,
it is customary in regression analysis to
automatically as a result of the regressio[illegible]

if doen expresses direct causation, coreferentiality should favour the use of doen (you cannot get more direct)

The internal factors

- syntactic construction type
- coreferentiality between matrix subject and infinitival subject/object
- animacy of matrix subject
- lexical colle[illegible]tional strength
- conceptual collocat[illegible]ength

PS why these ? for theoretical reasons,
it is customary in regression analysis to
automatically as a result of the regressio

if doen expresses direct causation, you expect more doen with animate matrix subjects (animate subjects have more control over the flow of energy)

The internal factors

- syntactic construction type
- coreferentiality between matrix subject and infinitival subject/object
- animacy of matrix subject
- lexical collocational strength
- conceptual co…tional strength

PS why these ? for theoretical reasons
it is customary in regression analysis t
automatically as a result of the regress

if the relevant factors are purely semantic ones (a model of causation), you don't expect any collocational idiomatization (lexical fixation)

Construction type / Coreferentiality

	- coreferentiality	+ coreferentiality
x doet/laat y_{subj} $V_{intrans}$	ik ... iets vallen	ik ... mij vallen
x doet/laat y_{subj} V_{trans}	ik ... hem doen	ik ... mij doen
x doet/laat z_{obj} V_{trans}	ik ... iets zien	ik ... mij verrassen
x doet/laat y_{subj} z_{obj} V_{trans}	ik ... iemand iets zien	ik ... iemand mij verrassen
x doet/laat z_{subj} door y_{pp} V_{trans}	ik ... de boom door hem vellen	ik … mij door iemand verrassen

Construction type / Coreferentiality

discarded cases:

- verbs that do not pattern independently: laten betijen
- optatives: laat ons hopen
- nominalizations: het laten varen van alle hoop
- grammaticalized idiomatic expressions: laat ons zeggen, laat staan dat

(and, of course, straightforward spurious hits)

Collocational measures

general introduction:

N	node: the element of interest
C	collocates: words within a certain span of N
C N	number of occurrences of C as collocate of N
~C N	occurrences of ~C (all other words) as collocate of N
C ~N	occurrences of C as collocate of ~N
~C ~N	occurrences of ~C as collocate of ~N

Collocational measures

the ratio C N / ~C N quantifies the popularity of C as a collocate of N

the ratio C ~N / ~C ~N quantifies the popularity of C as a collocate of all other nodes apart from N

comparing the two ratios tells you whether C is more typical as a collocate of N than as a collocate of any other node

specific statistics used here: log likelihood ratio

Collocational measures

general schema:

	+ N	~ N
+ C	+ C + N	+ C ~ N
~ C	~ C + N	~ C ~ N

starting with N as either doen or laten, the general schema can be filled out in several ways (through the selection of search domains and the selection of contrast sets)

Collocational measures

lexical collocation: how typical is a given verb as a collocate of doen (in either Flemish or Dutch)? and analogously for laten ?

e.g.

Dutch

+ V + doen	+ V ~ doen
~ V + doen	~ V ~ doen

where + V ~ doen = any other occurrence of V within CGN

Collocational measures

lexical distinctness: how typical is a given verb as a collocate of doen in comparison with laten (in either Flemish or Dutch)?

e.g.

Dutch

+ V + doen	+ V ~ doen
~ V + doen	~ V ~ doen

where ~ doen = + laten

Collocational measures

conceptual collocation: how typical is a given verb as a collocate of either doen or laten, i.e. how typical is it for causative construction?

e.g.

Dutch & Flemish

+ V + doen/laten	+ V ~ doen/laten
~ V + doen/laten	~ V ~ doen/laten

where + V ~ doen/laten = any other occurrence of V within CGN

Collocational measures

some context:

lexical collocation: the traditional form of collocational analysis, popularized within CL circles as 'collostructional analysis'

lexical distinctness: introduced by Gries/Stefanowitsch as 'distinctive collexeme' analysis - not used here, statistically less reliable than lexical collocation

conceptual collocation: a novel type of collocation analysis

Regression analysis

intro: what it does, what we have to look at, which regressions we'll consider

1 what it does:

construct a model explaining the variation in the data (in our case: the choice between doen and laten),

by stepwise adding the factors (as coded in the database) that contribute most to the reduction of the variation

Regression analysis

2 what we will have a look at:

what are the factors that are retained in the model ?

what is the predictive accuracy of the model ?

in what direction do the factors work (for or against doen/laten) ?

in what order are they added to the model ?

what are the significant values of the factors ?

Regression analysis

3 which regressions we will consider:

a) the dataset as originally coded, for the material as a whole

b) as in a), but separately for Flemish and Dutch

(we also looked at recoded datasets, but these results will not be presented here; they confirm the initial analysis)

(also: no attention to statistical interaction of factors in this presentation)

Global logistic regression

3975 observations, of which less than 10% doen

relevant factors, in order of importance:

- construction type

in contrast with the intransitive condition, transitives boost the presence of laten

Global logistic regression

- animacy: inanimate matrix subjects massively support doen, e.g. de wind deed hem huiveren

- country: Flemish has more doen than Dutch

- register: the majority of non-spontaneous, prepared text types significantly support doen

Global logistic regression

- collocational measures

a) significant conceptual collocation enhances laten: the more a verb is typically used in a causative construction, the more laten is used, i.e. laten is thé default verb for causatives

b) significant lexical collocation enhances doen: some verbs typically associate with doen (more than inanimacy of subject etc. predict); as a marked form, doen tends to be a lexical exception

Global logistic regression

- predictive accuracy

e.g. if the overall distribution is 90% laten, 10% doen, how much can you gain on the basis of the regression model ?

accuracy of dummy model is 92%

best accuracy of fitted model is 95%

→ a very strong and reliable model

Regional logistic regressions

- overall models are the same, and so is the order of inclusion of the factors: V1 ~ constr + anim + comp + sig.sem.col + sig.lex.col,
i.e. the difference is one of degree rather than principle

- no marked differences within register factor
(unlike many other typically Flemish forms, doen is not a marker of informality)

Regional logistic regressions

- the effect of the collocation measures is more outspoken in Dutch than in Flemish (see the estimates and the odds ratio's):

the existence of either a lexical or a conceptual association is more extreme in Dutch

this may be an indication of a more stabilised linguistic situation

Summary

- the default form for causatives is laten, to the extent that the more typically causative a construction is, the more readily it uses laten
- doen is a marked form, triggered by constructional (inanimacy of matrix subject, intransitivity of verb) ànd lexical factors
- doen is more formal, given its distribution over registers, than laten
- the restrictions on the use of doen are less outspoken in Flemish than in Dutch

Summary

is it possible to find a unifying interpretation for these results ?

the direct/indirect causation model is not completely adequate:

a majority of the predictions that we started off with is not confirmed

Summary

- intransitivity halts laten: not correct
- intransitivity boosts doen: correct
- coreferentiality boosts doen: not correct
- animacy boosts doen: not correct
- idiomaticity plays no role: not correct
- lectal effects are not expected: not correct

Summary

→ an alternative interpretative hypothesis:

doen is an archaic form; this ties in with all the relevant observations, i.e.

a) that it is typical for more formal registers

b) that it is sensitive to lexical associations (idiomatic effects as a form of relics)

c) that it occurs more in Flemish (which is known to be the more archaic variety in a number of respects)

Summary

d) that, semantically speaking, it seems to be retracting to one core form of causation, i.e. direct material causation

(the directness explains the intransitivity effect: transitives involve an intermediate entity)

(the material aspect relates to the inanimacy of the subjects, as opposed to the volitional causation of human subjects)

Refinements

further steps to take:

- expanding the descriptive basis
- refining the coding schema
- finetuning the collocational measures

Refinements

further steps to take:

- expanding the descriptive basis
- refining the coding schema
- finetuning the collocational measures

- the present data set is skewed towards spoken language, while at the same time, written language (component o) is significantly different
- there is an effect of register, but some registers (CGN components) are underrepresented

→ expand the database

Refinements

further steps to take:

- expanding the descriptive basis
- refining the coding schema
- finetuning the collocational measures

1 refining the animacy parameter

animals have now been coded as animate, but do they behave differently than people, i.e. do we need a cline from human to inanimate?

and what would be the position of human collectivities on the cline (from the team over the government to the nation)?

Refinements

further steps to take:

- expanding the descriptive basis
- refining the coding schema
- finetuning the collocational measures

2 refining the argument description

so far, only the matrix subject has been semantically classified, but what about the other arguments?

e.g. de bloemen laten hun kopjes hangen, 'the flowers let their heads hang low' – intransitive, inanimate subject, but doen seems unlikely

Refinements

further steps to take:

- expanding the descriptive basis
- refining the coding schema
- finetuning the collocational measures

3 refining the verbal classification

do semantic predicate classes (verbs of cognition, perception etc.) play a role?

should we have a separate code for verbs that are potentially ambiguous between a transitive and an intransitive reading (let the potatoes boil)?

does French have an influence on the Flemish data?

Refinements

further steps to take:

- expanding the descriptive basis
- refining the coding schema
- finetuning the collocational measures

how can we detect multiword collocational patterns, beyond the Aux Verb slot ?

Conclusions

1 bottom-up analyses pay off: schematic and only vaguely sunstantiated hypotheses of the direct/indirect causation type have to be refined when you look carefully at the data

→ even in such 'vague' areas as semantics and pragmatics, linguistics is an empirical science based on statistical hypothesis-testing

Conclusions

2 lectal variation plays an important role in the choice for doen or laten, both in terms of region (Dutch vs. Flemish)

and in terms of register (spontaneous vs. prepared sources)

→ we need to develop a lectally enriched mutivariate usage-based grammar