Journal of Language Modelling

VOLUME 6 ISSUE 1
JUNE 2018

Institute of Computer Science
Polish Academy of Sciences
Warsaw

Journal of Language Modelling

VOLUME 6 ISSUE 1
JUNE 2018

Articles

JOURNAL OF
LANGUAGE MODELLING

ISSN 2299-8470 (electronic version)
ISSN 2299-856X (printed version)
http://jlm.ipipan.waw.pl/

Published by IPI PAN
Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland

Circulation: 100 + print on demand

Layout designed by Adam Twardoch.
Typeset in XeLaTeX using the typefaces: *Playfair Display*
by Claus Eggers Sørensen, *Charis SIL* by SIL International,
JLM monogram by Łukasz Dziedzic.

EDITORIAL BOARD

Second order inference in natural language semantics

Stephen Pulman
Department of Computer Science, Oxford University

ABSTRACT

In this paper I look at a number of apparently trivial valid inferences (as well as some invalid and missing inferences) associated with the possessive construction and with different types of adjectival modification of nouns. In the case of possessives, all analyses I know of, whether implemented or not, systematically sanction invalid inferences. In the case of adjectives, there are some model-theoretic linguistic analyses that are adequate at a theoretical level, but no satisfactory practical computational implementations that I am aware of which capture the correct inference patterns.

A common thread between the possessive and the adjectival constructions is that to derive the correct inferences we need second order quantification. This is an uncontroversial move within model-theoretic formal semantics but a problem for computational semantics, since we have no fully automated theorem provers for anything other than first order logic (and only for subsets of first order logic do we have provers that are both fully decidable and efficient). I explore what is needed to provide a proof-theoretic account of the relevant inference patterns, and suggest some analyses requiring second order axioms. In order to make this a practical computational possibility I go on to propose two techniques for approximating such inferences in a first order setting. The suggested analyses have been fully implemented, and in an appendix I provide a small FraCaS-like corpus of relevant examples, all of which are handled correctly by the implementation.

Keywords: first order, second order, inference, adjectives, possessives

1 INTRODUCTION

The aim of this paper is to be able to capture some apparently trivial natural language inferences (and lack of inferences) involving adjective modification and possessive determiners, which like many other constructions turn out to have the property that second order quantification is required to capture these inferences. I will assume a simple and standard setting in which to address this problem, assuming that we have a syntax-driven compositional semantics producing logical forms for a (disambiguated) parsed sentence in a familiar way. These logical forms will ideally be sent to an automated theorem prover of some type (resolution, tableau...) which can mechanically check the validity of the inferences. A common version of this setting is to have the translations of declarative sentences or statements added as 'axioms' or 'premises', and then to have questions corresponding to the inferences we are interested in treated as 'theorems' to be proved, as for example in versions of the FraCaS inference suite (Cooper *et al.* (1996), MacCartney and Manning (2008)). The questions can be yes/no type questions where we will expect the answer 'yes' if there is a proof and either 'no' or 'don't know' otherwise (there's more to be said here: failure to find a proof does not always mean a negative answer), or Wh-questions where if there is a proof we will ideally return unifying substitutions corresponding to the values of the 'wh' constituent in the question.

Here is a very simple example:

All bankers are rich.	axiom: $\forall x.banker(x) \rightarrow rich(x)$
Jones is a banker.	axiom: $banker(jones)$
Is Jones rich?	prove: $rich(jones)$
Who is rich?	prove: $\exists x.rich(x)$

The first order logical (FOL) forms in the right hand column can be submitted to a first order theorem prover such as Prover9 (McCune (2005–2010)) and the answers retrieved (after some housekeeping) should be 'Yes' and 'Jones' respectively.

My modest aim in this paper is to be able to do something similar with inferences such as those described in the following sections, involving different types of adjectives, possessive determiners, and their combinations.

2 ADJECTIVES

For completeness, we will go through the standard examples of the inferential phenomena we are interested in even though, at least for adjectives, they are comparatively well-known. We begin with the simplest class of adjectives, usually called 'intersective', which sanction inferences like these:

(1) a. Jones is a red-haired farmer.
b. ⊨ Jones is red-haired.
c. ⊨ Jones is a farmer.

If we now add some extra information about farmers we get the following inference pattern:

(2) a. All farmers are gamblers.
b. ⊨ Jones is a red-haired gambler.

With what are commonly called 'gradable' or 'subsective' adjectives, we get a different pattern of inferences:

(3) a. Minnie is a large mouse.
b. ⊨ Minnie is a mouse.
c. ⊭ Minnie is large. *(can be valid with some contextual assumptions)*

(4) a. All mice are animals.
b. ⊨ Minnie is an animal.
c. ⊭ Minnie is a large animal.

(5) a. All mice are small animals.
b. ⊨ Minnie is a small animal.

Gradable adjectives have some implicit scale of comparison associated with them, and thus something can have contradictory properties if these are associated with different comparison scales. You can be a tall person but a short basketball player, for example. For many such adjectives it sounds odd to have the property ascribed unless the comparison scale is obvious from the context. In a few cases there can be a default comparison class, e.g. 'Mary is good' can be meaningful without a specific hidden parameter since for most people a

default parameter for 'good' will be 'behaviour' or 'character'. Some gradable adjectives like 'clever' or 'generous' have a further dimension, in that someone might be generous not only by comparison with other members of a class, but generous with respect to some properties (e.g. money) and not others (e.g. time). Recovering these relevant contextual parameters is a long way beyond the state of the art computationally, and so here we only use examples where the parameter is supplied linguistically, for example 'John is a tall man', and 'This is a red apple', rather than 'John is tall' or 'This apple is red'. Clearly what is tall for a man is not what is tall for a tree, and red for an apple is very different from what is red for a face, and until we know what this parameter is, few inferences are sanctioned.

A third class of adjectives are sometimes called 'privative', and whereas the first two classes have the property that from 'X is Adj Noun' we can always infer 'X is Noun', privatives do not behave in this way:

(6) a. Tony Blair is the former Prime Minister.
b. $\not\models$ Tony Blair is the Prime Minister.

(7) a. Smith showed an apparent proof of the theorem.
b. $\not\models$ Smith showed a proof of the theorem.

(8) a. He owns a fake diamond.
b. $\not\models$ He owns a diamond.

All of these adjectives have the property that from 'X is Adj Noun' the inference 'X is Noun' does not hold, and some have argued that for some cases, like 'fake', the inference to 'not-Noun' holds:

(9) a. This is a fake diamond.
b. $\models$ This is not a diamond.

Intuitions vary about this: Partee (2007) thinks that 'former P' entails 'not P now', whereas for me a sentence like 'In 2014, Obama was both the former and the current US President' is not contradictory.

Inferences from the complement of a privative adjective seem quite varied: 'This is a fake Picasso painting' does not entail 'This is a fake painting', whereas 'Bush is a former US president' does entail 'Bush is a former president'. However, 'Jane is a former fussy eater',

does not entail 'Jane is a former eater'.[1] Clearly there is more to be learned about such examples: for an interesting extended discussion and analysis of different types of privative adjectives, see Del Pinal (2015).

For some adjectives of this type, there are also some further interesting properties when combined with possessive determiners, such as the ambiguity of 'Mary's former mansion', which can be interpreted as referring either to the mansion that Mary used to own, or the building that Mary still owns which used to be a mansion. See Partee (2007) for discussion.

It is reasonably easy to specify truth conditions for **intersective** adjectives as follows, where **D**(x) = 'denotation of x':

> 'Jones is a red-haired farmer' is true iff
> **D**(jones) ∈ **D**(red-haired) ∩ **D**(farmer).

However, extending this definition to the other two classes requires appeal to notions which are not all intuitively clear and not very easy to pin down with mathematical precision. For **subsective** adjectives, perhaps:

> 'Minnie is a large mouse' is true iff **D**(minnie) ∈ {X | X a mouse larger than the relevant standard for mice}

Making the notion of "relevant standard" precise might involve assuming an ordering over mice by size (presumably adjusting for age) and fixing an interval representing the expected norm. As many people have commented (e.g. Kamp (1975)), this seems a little odd, in that it implicitly uses the comparative form of the adjective to define the semantics of the non-comparative form, whereas pre-theoretically one might have expected things to be the other way round.

In the case of **privative** adjectives, truth conditions seem to vary according to the specific adjective. For example, 'X is a former Y' is true iff **D**(X) ∈ **D**(Y at earlier time), and 'X is an alleged Y' is true (according to Morzycki (2014)) iff X is a Y in every possible world compatible with the allegation (although wouldn't 'X is an alleged Y iff someone has alleged that X is a Y' be simpler?).

[1] Thanks to a referee for this example.

It is clear that there is a large contextual component in the interpretation of all of these adjectives, and it is perfectly reasonable to pursue a style of analysis in which the logical form associated with them is rather minimal, most of the hard work being done by setting of various contextual parameters, with the analysis perhaps also involving probabilities or utilities (Rett (2014); Lassiter and Goodman (2017)). But whatever the undoubted merits of these approaches to defining interpretation conditions for adjectives or other context dependent constructs, the exercise is not very practically relevant for computational purposes, for which we need an explicit logical form that will support the relevant inferences proof theoretically, or which will lend itself to computationally tractable model building and checking techniques. In this respect, computational semantics for natural language is a rather different pursuit than purely linguistic semantics.

When constructing logical forms, if we are to be as compositional as possible, then any differences in the logical form of these three types of adjective under discussion must come either from some syntactic differences between the sentences in which they occur, or from their lexical properties. Since there seems to be no compelling evidence of a syntactic difference between these types of adjective (there are distributional differences to do with attributive and predicative uses of adjectives but this seems to cross-cut the present set of distinctions) I propose to build semantic differences into their lexical logical forms directly.

I will illustrate this with a small but precise fragment: a context-free grammar with associated semantic rules which build the meaning of a mother constituent by combining the meanings of daughter constituents. The meanings are expressed in a simply typed higher order logic of a familiar kind. For example, the first rule says that a Sentence (S) consists of a Noun Phrase (NP) followed by a Verb Phrase (VP) and that the meaning of the sentence is obtained by substituting the meanings of the NP and VP in the typed[2] higher order logic schema following the rule, which applies the NP meaning to the VP meaning.

[2] A type like `(et)t` is equivalent to (e → t) → t or `<<e,t>,t>` in other notations.

S	→	NP VP	: $NP_{(et)t}(VP_{et})$
NP	→	Det N'	: $Det_{(et)(et)t}(N'_{et})$
NP	→	Name	: $\lambda P_{et}.P(Name_{e})$
N'	→	N	: N_{et}
N'	→	Adj N'	: $Adj_{(et)(et)}(N'_{et})$
Adj	→	wooden, etc.	: $\lambda P_{et}x_{e}.wooden_{et}(x) \wedge P(x)$
Adj	→	small, etc.	: $\lambda P_{et}x_{e}.small_{e(et)t}(x,P)$
Adj	→	apparent, etc.	: $\lambda P_{et}x_{e}.apparent_{e(et)t}(x,P)$
Det	→	some, etc.	: $\lambda P_{et}Q_{et}.\exists x_{e}.P(x) \wedge Q(x)$ etc.
VP	→	Vi	: Vi_{et}
VP	→	Vt NP	: $Vt_{((et)t)et}(NP_{(et)t})$
Vi	→	snores, won, etc.	: $snore_{et}$
Vt	→	hits, is, likes, etc.	: $\lambda O_{(et)t}x_{e}.O(\lambda y.hit_{eet}(x,y))$

This grammar will deliver logical forms for our representative cases as follows, with some simplifications concerning the copula:[3]

Jones is a red-haired rugby player.
red-haired(jones) ∧ rugby-player(jones)
Minnie is a large mouse.
large(minnie, mouse)
Tony Blair is a former Prime Minister.
former(tonyblair,prime-minister)

For the intersective cases, we get the inferences we want immediately, since we have built conjunction into the lexical entry. For the subsective cases, we supply the non-logical constant encoding the adjective with an extra second order argument, which picks up the denotation of the noun as supplying the relevant comparison class. We can read small(x,P) as 'small by the standards relevant for P'. Of course, there is one crucial inference not proof-theoretically sanctioned by this logical form: to get the inference from adj(x,P) that P(x) we add (in an implementation, via an axiom schema) an axiom for each such adjective:

(10) $\forall xP.adj(x,P) \rightarrow P(x)$

[3] Following Montague, 'be' is translated as a transitive verb meaning '=', and the resulting logical forms can be simplified using the equivalence: $\exists x.P(x) \wedge x=a \equiv P(a)$. We notate this below as '$\Rightarrow_{=}$'.

We could instead have had a lexical entry for these adjectives that builds the inference in directly:

(11) $\lambda Px.small_{e(et)t}(x,P) \wedge P(x)$

but this is not in my view particularly compositional.[4] While compositionality is difficult to define (see the survey and discussion in Szabó (2017)) and may be no more than a methodological rule of hygiene, some simple principles would surely include a requirement that a single content word in a sentence should correspond to no more than one component of a logical form (whereas function words in most frameworks have to be allowed to introduce an amount of logical 'glue').

In order to cope with the privative cases we simply refrain from generating these axioms and so we (correctly) cannot infer from apparent(x,P) that P(x). For those privative adjectives, if there are any, that sanction the negation of the property we can add an axiom:

(12) $\forall xP.adj(x,P) \rightarrow \neg P(x)$

This is all very tidy and makes it easy to define truth conditions for these logical forms with rather less contextual clutter than would be needed for simpler forms that did not include these parameters. But these logical forms do not solve our computational inferential problem because they involve second order arguments to predicates. The state of the art in automated inference is that we have reasonably efficient general purpose theorem provers for first order logic (with equality) except that they are bounded by the inescapable semi-decidability of FOL, and the unpredictable computational complexity of general inferences. By restricting the expressivity of FOL to tractable subsets (Baader *et al.* (2003)) we can guarantee good performance, but only for a small number of the cases we would like to handle.

Regrettably, it is both in theory and in practice impossible to reason directly, as we would like to do, with higher order logics: even the notion of higher order unification needed as a component is undecidable (Huet (1975)). There do exist some higher order logic proof assistants like HOL (`https://hol-theorem-prover.org/`), Isabelle (`https://www.cl.cam.ac.uk/research/hvg/Isabelle/`), and Coq

[4] The same objection, and a version of the same solution, are relevant to the treatment of intersective adjectives just given.

(`https://coq.inria.fr/`). However, these are not fully automatic theorem provers, but interactive systems requiring human guidance and input at every step, and are usually used for checking already generated proofs. It is, however, possible to write special purpose proof 'tactics' to guide a proof assistant like Coq to carry out some specific higher order logic inferences derived from natural language expressions semi-automatically, and in a series of papers from Chatzikyriakidis and Luo (2014) onwards, Chatzikyriakidis and Luo have carried out such experiments on a variety of constructions. Similar efforts are described in Mineshima *et al.* (2015) and related papers. However, while this is an interesting experiment from the point of view of validating particular higher order analyses of linguistic phenomena, it is important to recognise that it is a very different exercise from our current aims: the approach is not a general purpose technique of the type we would like, but something which will only work on prespecified patterns and derivations. The results could not, for example, form a component of an automatic natural language processing system performing these inferences as part of an application task like question answering or task-oriented dialogue.

The limitations of automated higher order logic inference constitute a real barrier to computational semantics of natural language, because like the analysis here, many natural language constructs are intrinsically higher order. Some obvious ones are generalised quantifiers and intensifying modifiers, where outline logical forms are shown below. 'Most' will be a function from a noun meaning to a function from verb meanings to truth values. 'Very' will be a function from adjective meanings to adjective meanings.

(13) a. Most dogs bark. = (most dog) bark

b. John is very tall. = (very (tall)) john

One approach to this problem, since the specialised HOL or Coq proof tactics just mentioned are not a general solution, is to try to translate or compile the higher order forms to something that a FOL prover can deal with. There are a number of strategies that have been tried: reification or 'ontological promiscuity' attempts to 'compile out' the higher order aspects by adding different types of abstract individuals to first order models. Some common examples of this strategy in-

clude event analyses of verb modification (Davidson (1967)), although in this case, arguably, there is also some linguistic motivation:

(14) a. John ran in the park. = (in the park)(run)(john)
b. ⇒ ∃e.run(e,john) ∧ in(e,the park)

or the so-called 'standard translation' of modal logic:

(15) □p ⇒ ∀w.R(thisWorld,w) → p(w)

which translates 'necessarily p' into 'in all worlds w in the appropriate relation R to this world, p is true in w'.

Hobbs (1985) has been a notable advocate of this approach, and it has been applied to the semantics of adjectives in Amoia and Gardent (2007). But it's not obvious how such a strategy could help us here, in the general case at least, although it has been used successfully in specific limited domains where we can precompute values for the various adjective parameters. Let's assume we try to eliminate the second order arguments in our subsective Adj meanings by adding entities representing standards of Adj-ness for those adjectives. We will then translate 'John is a tall man' as something like:

(16) ∃s.tall(john,s) ∧ man(john) ∧ tallness-for-men(s)

'John is tall to s, where s is that degree of tallness for men which qualifies as being tall'. (Note that in forms like 'John is tall' we will have to fill in the relevant noun parameter from context or non-linguistic knowledge, but this is the case for all approaches). So far, so good: it is easy to see how to make implementational sense out of this, given a sufficiently well structured domain. However, when we look at what else we need to do to make this analysis work things get more complicated: for example, we need to ensure that an adjective interacts properly with related (usually antonymous) adjectives:

(17) John is a tall man. ⊨ John is not a short man.

∀xyz.tall(x,y) ∧ tallness-for-men(y) →
¬(short(x,z) ∧ shortness-for-men(z))

This is doable, if a little clumsy, and as we extend similar axioms we need to be careful to ensure that 'John is not short' does not wrongly entail 'John is tall'. A more serious problem is that there are

a potentially infinite number of such 'adjness-for-X' entities and their predicates, and therefore a potentially infinite number of such relatedness axioms. This happens because it is possible to combine adjective modification in principle to an arbitrary depth, essentially creating 'standards of comparison' on the fly:

(18) a. This is an old American building.
b. This is an older mid-period Anglo-Saxon religious site.

The interpretation we are interested in here is that on which each adjective modifies everything that follows it, rather than the usually possible 'conjunctive' reading on which each adjective just modifies the head noun. So we need a standard of comparison for age relevant for 'American building', which will be different from that for 'English building', as well as a standard for mid-period Anglo-Saxon religious sites. 'Mid-period' is itself subsective, the standard for that type of Anglo-Saxon religious sites will be different from that for Anglo-Saxon religious sites of all periods, and so on. The recursive nature of adjectival modification means that there is no limit in principle to the number of such standards and so we cannot just define them all in advance, nor can we list in advance all the required axioms connecting antonyms.

However, our second order analysis of these adjectives generalises quite cleanly to this case, without requiring separate axioms for each further combination:

(19) a. This is an old American building. =
b. old(this,λx.American(x) $\wedge$ building(x))

(20) a. This is an old mid-period Anglo-Saxon religious site. =
b. old(this,λx.mid-period(x, λy.Anglo-Saxon(y) $\wedge$ religious(y) $\wedge$ site(y)))

and the interaction with related predicates only needs one (second order) axiom (again generated from a schema, we assume), which quantifies over every possible standard of comparison:

(21) $\forall$xP. old(x,P) $\rightarrow$ $\neg$(young(x,P))

(22) $\forall$xP. tall(x,P) $\rightarrow$ $\neg$(short(x,P))

While this is satisfactory from the point of view of linguistic analysis, we are unfortunately still no nearer to a solution to the problem of how to automate inferences involving these logical forms: they are still second order.

3 POSSESSIVES

We turn now to possessive determiners, an apparently simple construction, but one which on closer inspection has several interesting properties. There are a number of relevant well-known properties of possessives for us to bear in mind when trying to uncover their inferential properties, as well as some less well-known properties. It is a striking fact, discussed further below, that that all of the well-known analyses of possessives sanction invalid inferences involving them.

Firstly, an obvious point to make is that the relation between possessor and possessed can vary and is not just restricted to a small set of semantic notions like 'ownership', 'part of', and the like; rather, it can depend on almost any feature of the linguistic or non-linguistic context:

The table's leg...	Monday's lecture...
America's invasion of Iraq...	John's measles...
John's dog...	John's brother...
John's portrait...	etc.

For example, 'John's dog' can mean the dog that John owns, the dog that John has just sold, the dog that John has just bet on to win in a race, etc. This wide contextual dependence, as with adjectives, makes it perfectly reasonable to adopt an analysis on which logical forms are relatively simple, and all the heavy lifting is done by setting of various contextual parameters. But as we argued when discussing adjectives, this is not a stance that is open to anyone wanting an implementable account of the inferences associated with such constructions.

Secondly, what we are calling the possessive comes in various syntactic forms:

(23) a. John's picture/team/sister
b. a picture/team/sister of John's
c. a picture/*team/sister of John
d. That picture/team/sister is John's

As the 'team' examples show, there are some acceptability variations associated with the difference between what are often called 'relational' and 'sortal' nouns. Relational nouns implicitly correspond to two-argument predicates whereas sortal nouns are more naturally modelled as one-argument predicates. As we will see later, this does not necessarily correspond to a syntactic distinction.

Third, relational and sortal nouns also seem to sanction inference patterns of differing acceptability (de Bruin and Scha (1988)), where we interpret 'has' in the following as denoting the same relation as the possessive:

(24) a. John's cars are wrecks. $\models$
b. Some wrecks of John's are cars; Some wrecks/cars are John's.
c. John has some wrecks; John has some cars.

(25) a. John's brothers are musicians. $\models$
b. ?Some musicians of John's are brothers.
c. ?Some musicians/brothers are John's.
d. ?John has some musicians; John has some brothers.

Despite these differences in acceptability, I would prefer not to distinguish relational vs. sortal nouns syntactically. This is because of the fourth observation: that all relational nouns can be interpreted as sortal in the right context, as many people have pointed out:

(26) The headmaster has difficulty dealing with his parents.

(Parents' evening context: headmaster is talking to parents of the children in his school.)

(27) John's famous wife is Victoria Beckham.

(John is one of several journalists tasked with writing a piece about famous men with equally famous wives.)

In the following, we will not attempt to capture all of the interesting properties of possessives in our analysis. Instead, we will concentrate on a quite modest ambition: we would like an implemented analysis of the possessive which allows us to avoid the invalid inferences of existing analyses, to be described later, and to capture valid inferences like the following:

(28) Smith is Jones's plumber. ⊨ Smith is a plumber.

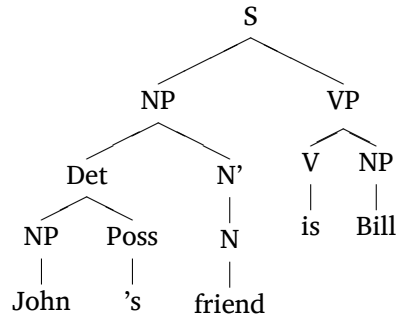
(29) a. John's old wooden toy broke. ⊨
b. John's wooden toy broke.
c. A wooden toy broke.
d. A toy broke.

(30) a. The student's essay's title intrigued Jones. ⊨
b. An essay's title intrigued Jones.
c. A title intrigued Jones.

(31) a. All John's brothers are rich.
b. Bill is John's brother.
c. ⊨ Bill is rich.

3.1 *An initial simple analysis*

A simple analysis (variants of which can be found in many places, for example Bos *et al.* (2004), or more recently, Steedman (2012)) takes the possessive morpheme *'s* (or just *'* for plurals) to be a function from NP meanings to Det meanings introducing an abstract two-place 'of' or 'poss' relation, usually assumed to be subject to further contextual resolution. In the illustrative framework we are using this would be implemented as follows:



(32) John's friend is Bill. = ∃x.friend(x) ∧ of(x, John) ∧ x=Bill

We need to add some rules to our earlier fragment to produce such an analysis:

Det → NP poss : $\text{poss}_{((et)t)(et)(et)t}(\text{NP}_{(et)t})$
poss → ' or 's : $\lambda O_{(et)t} P_{et} Q_{et}.O(\lambda y.\exists x.P(x) \wedge \text{of}_{eet}(x,y)) \wedge Q(y)$

In principle, we ought to be able to leave the 'of' predicate unresolved – not the least because this kind of contextually sensitive resolution is a completely unsolved computational inference problem – and still get most of the inferences we would like to get. But it turns out that this will lead us astray. If we leave 'of' unresolved, we will sanction some incorrect inferences:

A: John's brother is Bill. = $\exists x.\text{brother}(x) \wedge \text{of}(x, \text{John}) \wedge x=\text{Bill}$
$\Rightarrow_{=} \text{brother}(\text{Bill}) \wedge \text{of}(\text{Bill},\text{John})$

B: Bill is a doctor. = $\exists x.\text{doctor}(x) \wedge x=\text{Bill}$
$\Rightarrow_{=} \text{doctor}(\text{Bill})$

C: Bill is John's doctor. = $\exists x.\text{doctor}(x) \wedge \text{of}(x, \text{John}) \wedge x=\text{Bill}$
$\Rightarrow_{=} \text{doctor}(\text{Bill}) \wedge \text{of}(\text{Bill},\text{John})$

Now C is provable from the conjunction of A and B, incorrectly; whereas C is not a valid inference from A and B.

Perhaps it was a mistake to leave 'of' unresolved? 'Of' can be contextually interpreted as 'has', 'owns', or as an arbitrarily complex context-dependent relation like 'bet-on-by', or as the relation associated with a relational noun, if present:

(33) a. John's dog won. =
b. $\exists x.\text{dog}(x) \wedge \text{owned-by}(x, \text{John}) \wedge \text{won}(x)$
c. $\exists x.\text{dog}(x) \wedge \text{bet-on-by}(x, \text{John}) \wedge \text{won}(x)$
d. etc.

(34) a. John's brother arrived. =
b. $\exists x.\text{brother}(x) \wedge \text{brother-of}(x,\text{John}) \wedge \text{arrived}(x)$

If we now interpret 'of' in A above as the two-place relation 'brother(Bill,John)', and as something else in C (for example, 'treated-by'), then the incorrect inference will not be made.

Unfortunately, contextual interpretation doesn't always solve this problem. Although our invalid inference will not go through when relational nouns are involved (at least if we use them as the source for the contextually dependent resolution option) we cannot always

guarantee validity for examples involving sortal nouns. Consider the following example:

A: Smith is Bill's plumber. = (interpret 'of' as 'works-for')
$\Rightarrow_{=}$ plumber(Smith) $\wedge$ works-for(Smith,Bill)

B: Smith is also a decorator.
$\Rightarrow_{=}$ decorator(Smith)

C: Smith is Bill's decorator.
$\Rightarrow_{=}$ decorator(Smith) $\wedge$ works-for(Smith, Bill)

It's surely impossible to argue that 'of', interpreted as a contextually dependent 'works-for', or 'employed-by' relation, should be instantiated differently in A and C, and under these interpretations the unwanted inference will still go through. The analogous bad inference will also go through even where we do have a relational noun but where it is interpreted sortally. If we interpret the possessive as something like 'taught by' in:

A: The noisy class were Mr Smith's children.

B: The noisy class are also the Latin class.

C: The noisy class are Mr Smith's Latin class.

then C should not follow from A and B, but it will do so given the logical forms assigned by this analysis, even after resolution.

3.2 *Two further, more sophisticated, analyses*

In Partee and Borschev's analysis (Partee and Borschev (2003)), the possessive morpheme introduces a lot more structure:

(35) John's = $\lambda N.\lambda P.\ \exists x.[Sort(N)](x) \wedge R_{gen}(x,John) \wedge P(x)$

In their analysis, relational and sortal nouns are assigned to different types: for example, $brother_{eet}$ and $team_{et}$. In order to keep the types straight in composition they define a 'typeshifting' operator, 'Sort', defined thus:

(36) a. A: $Sort(N_{eet}) = \lambda x.\exists y.N(x,y)$

b. B: $Sort(N_{et}) = N$

Applying clause A of the definition to a relational noun like $brother_{eet}$ = $\lambda x.\lambda y.brother_2(x,y)$ produces a one-argument version with the same type as the corresponding sortal noun: $brother_1 = \lambda x.\exists y.brother_2(x,y)$.

The relation "R_{gen}" is their version of our 'of' relation, to be contextually interpreted, but with a default preference for the relational version of a noun N_2 if N_1 is present. This approach gives analyses like the following:

(37) a. John's team won. =

b. $[[\lambda N.\lambda P.\exists x.(Sort(N))(x) \wedge R_{gen}(x,John) \wedge P(x)](\lambda y.team(y))](won)$

c. $\Rightarrow_{\beta} \exists x.team(x) \wedge R_{gen}(x,John) \wedge won(x)$

The relation 'R_{gen}' can then be contextually interpreted as something like 'played-in-by' or 'supported-by', as appropriate. For the relational case:

(38) a. John's brother arrived. =

b. $[[\lambda N.\lambda P.\exists x.(Sort(N))(x) \wedge R_{gen}(x,John) \wedge P(x)](\lambda y.\lambda z.brother_2(y,z))](arrived)$

c. $\Rightarrow_{\beta} \exists x.brother_1(x) \wedge R_{gen}(x,John) \wedge arrived(x)$

then R_{gen} can be instantiated to the original relational version of 'brother':

(39) $\exists x.brother_1(x) \wedge brother_2(x,John) \wedge arrived(x)$

I do not find this a particularly satisfying or elegant analysis. Note that for sortal nouns, this is just a variant of our first simple analysis, and so it will also sanction the same set of invalid inferences. In the case of relational nouns, the treatment is surely very clumsy. Initially, the contextually appropriate two-place relation is accounted for in the analysis, but transformed to a 1-place relation to keep the types straight. Then the original two-place relation has to be recovered again by inference. Furthermore, the strategy of giving relational nouns a different type from sortal nouns means that everything that can combine with N (Det, Adj, etc.) will now have to be polymorphic, i.e. set up to expect two different types: `eet` and `et`, or alternatively have the 'Sort' operator wrapped around it to coerce two-argument predicates to one-argument predicates. This seems a high price to pay, both linguistically and computationally.

An influential alternative analysis by Peters and Westerståhl (2006) (see also Peters and Westerståhl (2013)) makes several perceptive contributions to our understanding of possessives. In their

analysis, possessives involve two quantifiers, one associated with the NP in the possessive Det phrase, and the other either explicit, as in:

(40) Several of each farmer's sheep are infected.

or contextually inferred:

(41) a. John's fingers are clean. (all of them)
b. John's fingers are dirty. (just some)

A second concern in their analysis is to capture the phenomenon of 'narrowing', as in:

(42) a. Most people's grandchildren like them.
b. Many planets' moons are visible.

In these examples, 'most/many' are quantifying over 'people with grandchildren' or 'planets with moons' rather than just 'people' or 'moons'. A related phenomenon is discussed by Bos (2009) noting that possessives involving superlatives:

(43) a. London's most expensive restaurant...
b. Milan's best player...

require a comparison set that involves the possessor as well as the possessed.

Peters and Westerståhl give truth conditions for two variants of the possessive construction (their account is couched in model theoretic terms), with or without an explicit quantifier (Q_2) in a predeterminer position:

(44) a. Q_1 C's As are B
b. Q_2 of Q_1 C's As are B

Peters and Westerståhl define a 'Poss' higher order operator (distinct from the "poss" morpheme) which has four arguments: (i) an explicit (sometimes implicit) quantifier in the possessive determiner phrase, (ii) the explicit or contextually inferred predeterminer quantifier, (iii) the possessed nominal relation, and (iv) a two-place relation 'R' corresponding to our 'of' and also a placeholder for a contextually inferred relation. They further define, for a two-place relation R and a set A:

(45) a. R_a = {b : R(a,b)} or in our logical form notation λb.R(a,b)

b. $dom_A(R)$ = {a: A ∩ $R_a \neq \emptyset$} or λa.∃b.A(b) ∧ R(a,b)

The expression in (a) denotes the set of things possessed by *a* and in (b) the set of objects that possess something in A. Now the truth conditions for an expression involving 'Poss' are defined as:

(46) $Poss(Q_1,C,Q_2,R)(A,B) = Q_1(C \cap dom_A(R), \{a: Q_2(A \cap R_a, B)\})$

Q_2, as above, is the inferred or predeterminer quantifier. Read this expression as:

(47) Q_1 C x that 'possess' an A are such that Q_2 A that x 'possesses' are B

It is assumed that even where Q_1 is not explicit, as in 'John's...' there is an implicit non-vacuous universal quantifier involved. The syntactic structure assumed for the case where there is an explicit predeterminer quantifier is illustrated in Figure 1. The case where the second quantifier is implicit is illustrated in Figure 2, and a concrete example of this phenomenon is offered in Figure 3.

In Figure 3, interpreting 'R' as 'own', we arrive at the interpretation:

(48) a. Poss(most,students,every,own)(cars,rusty)

b. = most(students ∩ dom_{car}(own), {a: every(car ∩ own_a, rusty)})

Translated to our logical form notation:

(49) most(λx.student(x) ∧ ∃b.car(b) ∧ own(x b),
λa.every(λy.car(y) ∧ own(a,y),rusty))

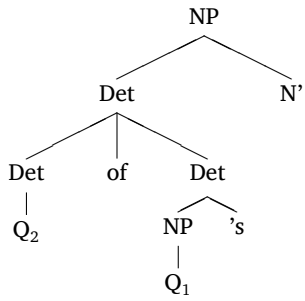


Figure 1:
Case with explicit predeterminer quantifier

Figure 2:
Case with implicit predeterminer quantifier

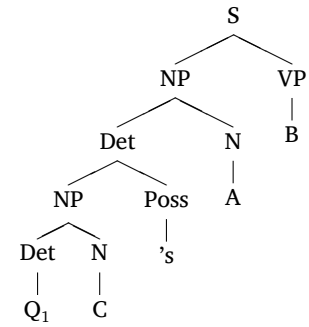
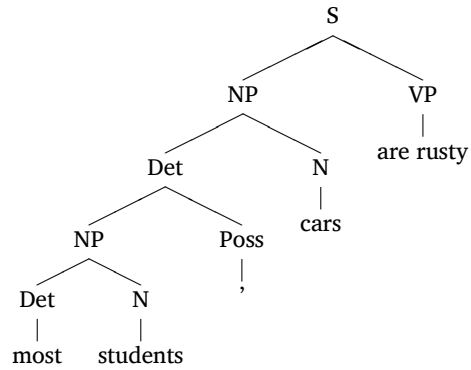


Figure 3:
An example of the phenomenon in Figure 2, where 'R' is interpreted as 'own'



While this is one of the most sophisticated analyses of the possessive in the literature, there are still a number of problems with it. For example, it is not clear that the 'implicit predeterminer quantifier' is specific to possessives or is an instance of the implicit quantification that is needed anyway for bare plural nouns (Lauri Carlson, p.c.). Furthermore, as the authors point out, this account is not fully compositional, since 'Poss' needs access to the components of its sister NP separately in order to build 'narrowing' into the truth conditions. In Peters and Westerståhl (2013) this is described as "second level" compositionality, accessing immediate constituents of immediate constituents, as opposed to "first level" compositionality. (We will ignore narrowing in what follows, for simplicity.)

For our purposes the most salient shortcoming of this analysis is that, however R is contextually interpreted, provided it is interpreted consistently, the unwanted inference in our 'plumber' and 'decorator' case will still go through on Peters and Westerståhl's analysis. This

is, as we shall see shortly, because any binary relation R is incapable of capturing the dependencies involved in blocking the invalid inferences.

Johan Bos (Bos (2009)) is aware of the problem we have signalled and suggests an analysis (that he attributes to Yuliya Lierler and Vladimir Lifschitz) in which we translate sentences like 'Vincent is Mia's husband' as:

(50) person(Vincent) ∧ ∃y.role(Vincent,y) ∧ husband(y) ∧ of(y,Mia)

paraphrased as 'Vincent is a person who is playing the role of Mia's husband'.

His suggestion is not sketched in full detail, and it may be possible to extend it to overcome these objections, but as it stands this analysis leaves much to be desired, in my view. To begin with, it is highly non-compositional: there are no words in the sentence corresponding to the logical form predicates 'role' and 'person'. Secondly, although the analysis certainly blocks our unwanted invalid inference, it also *fails* to sanction a basic and valid inference that we want to capture: If Vincent is Mia's husband, then Vincent is a husband: husband(Vincent). Thirdly, it is not clear how to extend the analysis to complex nominals: Bos's discussion suggests that we would get something like the following analysis:

(51) a. John's wooden toy disappeared. =

b. ∃x.thing(x) ∧ ∃y.role(x,y) ∧ wooden(x) ∧ toy(y) ∧ disappear(x)

We can successfully infer that 'Something wooden disappeared', but not that 'A toy disappeared', only that 'Something with the role of a toy disappeared'.

4 A HIGHER ORDER $of_{ee(et)t}$ RELATION

There is a relatively simple solution, linguistically at least, to this problem. Intuitively it is clear that what allows the invalid inferences to go through is that any binary possessive relation simply relates possessor and possessed, but does not capture in what respect the possessive relation holds. This respect is the property denoted by the N' constituent following the possessive 'NP's' determiner (which may have

to be recovered by ellipsis in some cases). If we make our possessive morpheme in our grammar fragment slightly more complex by giving it a semantics as follows:

(52) $\text{poss} = \lambda O_{(et)t} P_{et} Q_{et}.O(\lambda y.\exists x.P(x) \wedge \text{of}_{ee(et)t}(x,y,P)) \wedge Q(x)$

then we now make the respect in which the possessive relation holds an explicit argument of the 'of' placeholder relation. The "of" predicate is now of type ee(et)t: a function from individuals to individuals to properties to truth values. This is sufficient to block our unwanted inference:

(53) a. A: Smith is Bill's plumber.

b. $\Rightarrow_{=}$ plumber(Smith) $\wedge$ of(Smith,Bill,plumber)

(54) a. B: Smith is also a decorator.

b. $\Rightarrow_{=}$ decorator(Smith)

(55) a. C: Smith is Bill's decorator?

b. $\Rightarrow_{=}$ decorator(Smith) $\wedge$ of(Smith, Bill,decorator)

Now the unwanted inference does not go through. Note that this analysis is rather uncompositional, by the criteria we outlined earlier, in that one word corresponds to two identical non-logical constants in the logical form. We can make the analysis simpler and more compositional by dropping the repetition at the cost of two additional second order axioms:

(56) $\text{poss} = \lambda OPQ.\exists x.O(\lambda y.\text{of}(x,y,P)) \wedge Q(x)$

(57) Smith is Bill's plumber/brother. $\Rightarrow_{=}$ of(Smith,Bill,plumber/brother)

In order to recover the inference that Smith is a plumber, or where the noun is relational and interpreted relationally, as in 'a brother of Bill' we need these axiom schemata:

(58) a. A: $\forall xyP.\text{of}(x,y,P) \rightarrow P(x)$ ***(sortal N)***

b. B: $\forall xyP.\text{of}(x,y,P) \rightarrow P\text{-of}(x,y)$ ***(relational N)***

Note that we do not have to resolve 'of' to avoid bad inferences, and we do not need to distinguish sortal and relational N syntactically: these axioms capture their semantic differences.

Note also that unlike Bos's suggested analysis, ours generalises smoothly to complex nominal cases:

(59) a. John's wooden toy disappeared.
b. $\exists x.of(x,John,\lambda y.wooden(y) \wedge toy(y)) \wedge disappeared(x)$

Via axiom A we can deduce:

(60) $\exists x.[\lambda y.wooden(y) \wedge toy(y)](x) \wedge disappeared(x)$

and then by β-reduction and conjunction both that:

(61) a. A toy disappeared.
b. $\exists x.toy(x) \wedge disappeared(x)$

and that:

(62) a. Something wooden disappeared.
b. $\exists x.wooden(x) \wedge disappeared(x)$

Semantically it also follows from this last sentence that:

(63) John's toy disappeared: $\exists x.of(x,John,toy) \wedge disappeared(x)$

However, in order for us to be able to show this proof-theoretically we need something more elaborate. Intuitively, we want to be able to say that if $of(x,y,P)$ and P implies Q, then also $of(x,y,Q)$. If Fido is John's cat, and all cats are animals, then Fido is John's animal. Something like the following would suffice for this particular case, where the entailment involves conjunction, but later we will need something more general:

(64) $\forall xyPQ.\ of(x,y,\lambda z.P(x) \wedge Q(z)) \rightarrow of(x,y,P) \wedge of(x,y,Q)$

We will return to such cases below.

5 COMPUTATIONAL IMPLICATIONS

Our second-order analysis may be linguistically fine, but computationally it does not yet solve our problems. As already remarked, we cannot do second (or higher) order logic theorem proving automatically

except for some very special restricted cases, beyond which our analysis lies. So although we cannot hope for a fully general solution to the problem of automated inference for analyses like the ones developed so far using second order logical forms, in this section we will explore some heuristic techniques which may enable us to implement a special purpose first order solution, still bearing in mind that we only have computationally efficient reasoning for fragments of first order logic.

In this section we explore two alternatives, both of which try to make our second order reasoning look like first order proofs.

5.1 *Encoding via combinators*

As illustrative examples let us focus on some simple possessive inferences we want to capture:

(65) Bill is John's dentist $\models$ Bill is a dentist

Our earlier analysis, after equality simplications, will give these sentences the following logical forms:

(66) $\text{of(Bill,John,dentist)} \models \text{dentist(Bill)}$

A second slightly more complex example we would like to be able to handle is:

(67) a. John's wooden toy disappeared. $\models$
b. John's toy disappeared.
c. A toy disappeared.

(68) a. $\exists x.\text{of}(x,\text{John},\lambda y.(\text{wooden}(y) \wedge \text{toy}(y))) \wedge \text{disappeared}(x) \models$
b. $\exists x.\text{toy}(x) \wedge \text{of}(x,\text{John},\text{toy}) \wedge \text{disappeared}(x)$
c. $\exists x.\text{toy}(x) \wedge \text{disappeared}(x)$

We will assume the following axioms, introduced earlier:

Axiom A: $\forall xyP.\text{of}(x,y,P) \rightarrow P(x)$
Axiom B: $\forall xyPQ.\text{of}(x,y,\lambda z.P(z) \wedge Q(z)) \rightarrow \text{of}(x,y,P) \wedge \text{of}(x,y,Q)$

Notice that in the intended application of these axioms, applicability would be determined by higher order matching (which **is** decidable: Stirling (2010)) and thus would generalise to sequences of three or

more conjuncts inside the lambda term. However, our approximation will not behave in this way and so in reality we will need a version for 2, 3, etc. conjuncts.

The basic idea is to encode higher order terms as first order expressions via combinators, following Hurd (2002) who used this technique to automate some of the components of a human-assisted higher order proof. Our logical expressions are less general than those treated by Hurd, since only second order arguments are involved, and they are either single predicate constants, or lambda terms with complex terms formed by connectives, but no quantifiers, in their body. We can thus apply the usual first order normal form transformations needed for a resolution or tableau theorem prover to our logical forms to obtain clauses (disjunctions of literals), with the extra feature that any second order arguments like those we are using are 'frozen': their outermost lambda functor will be regarded as a function symbol and no transformations will take place inside that lambda term.

We now transform each literal. The first step is to represent literals in applicative form, using a two-argument functor 'a' meaning 'apply'. Since 'a' is a function symbol and not a predicate, to respect first order syntax and semantics we have to wrap a dummy predicate 'p' around the translation:

(69) a. sleep(john) = p(a(sleep,john))
b. like(john,jane) = p(a(a(like,john),jane))

Now we can represent predicate variables as ordinary first order variables, so that for example ∃P.P(j) = ∃P.p(a(P,j)), where the occurrences of P on the right hand side are first order.

Our axiom A, in implicational rather than clausal form, now looks like this:

(70) p(a(a(a(of,X),Y),Q)) → p(a(Q,X))

However, the more complex axiom B has the following form at this stage, still containing a lambda expression:

(71) p(a(a(a(of,X),Y),λZ.a(a(and,a(P,Z)),a(Q,Z)))) →
p(a(a(a(of,X),Y),Q)) ∧ p(a(a(a(of,X),Y),R))

We need to eliminate all lambda expressions, of course. It is well known that we can completely eliminate variables from a lambda-

Figure 4: Translation function T: eliminating variables

T[x]	⇒ x	
T[(E1 E2)]	⇒ (T[E1] T[E2])	
T[λx.E]	⇒ (**K** T[E])	(if x is not free in E)
T[λx.x]	⇒ **I**	
T[λx.λy.E]	⇒ T[λx.T[λy.E]]	(if x is free in E)
T[λx.(E1 E2)]	⇒ (**S** T[λx.E1] T[λx.E2])	(if x is free in both E1 & E2)
T[λx.(E1 E2)]	⇒ (**C** T[λx.E1] T[E2])	(if x is free in E1 but not E2)
T[λx.(E1 E2)]	⇒ (**B** T[E1] T[λx.E2])	(if x is free in E2 but not E1)
T[λx.(E x)]	⇒ T[E]	(if x is not free in E: this is eta reduction)

calculus based logic by using 'combinators'. We can use this fact to further try to squeeze our second order expression into something that can be handled by a first order prover. There are many variant formulations of variable-free combinator calculi, but we will use a familiar one, also used by Hurd:

Ix	=	x (identity)
Kxy	=	x (make constant function)
Sxyz	=	xz(yz) (generalised application)
Cfxy	=	fyx (special case of S)
Bfgx	=	f(gx) (special case of S)

For completeness, we give the usual definition of a translation function **T** that will eliminate lambda terms and their variables (Figure 4), where E1 and E2 are any well formed HOL expression.

Now our axiom B looks like this, in implicational form:

(72) p(a(a(a(of,X),Y),a(a(**S**,a(a(**B**,and),Q)),R))) →
p(a(a(a(of,X),Y),Q)) ∧ p(a(a(a(of,X),Y),R))

Given axiom B and the applicative logical form for 'Bill is John's dentist':

(73) of(Bill,John,dentist) = p(a(a(a(of,Bill),John),dentist))

it is (relatively!) easy to see that the applicative version of this logical form will (first order) unify with the antecedent of the implication in the applicative form of axiom B, with bindings X=Bill, Y=John, Q=dentist allowing us to deduce:

(74) p(a(dentist,Bill)) = dentist(Bill)

Perhaps less easy to see, as the applicative forms become less human readable, is that we can also make some of the deductions we wanted from:[5]

(75) a. John's wooden toy disappeared.
b. $\exists x.of(x,John,\lambda y.(wooden(y) \wedge toy(y))) \wedge disappeared(x)$

Using axiom B (in applicative form) we can deduce the equivalent of ...of(x,John,toy)... and from axiom A ...toy(x)... enabling us to prove the queries:

(76) a. $\exists x.toy(x) \wedge of(x,John,toy) \wedge disappeared(x)$
b. $\exists x.toy(x) \wedge disappeared(x)$

5.2 *Adjective inferences*

We can encode our adjective inferences in the same way:

(77) a. $\forall xP.\ small(x,P) \rightarrow P(x) \Rightarrow$
b. $p(a(a(small,X),P)) \rightarrow p(a(P,X))$

(78) a. $\forall xPQ.\ small(x,\lambda y.P(y) \wedge Q(y)) \rightarrow P(x) \wedge Q(x) \Rightarrow$
b. $p(a(a(small,X),a(a(\mathbf{S},a(a(\mathbf{B},and),a(a(\mathbf{B},P),\mathbf{I}))),a(a(\mathbf{B},Q),\mathbf{I}))))$
c. $\rightarrow a(P,X) \wedge a(Q,X)$

These axioms and others will enable us to capture inferences like:

(79) a. Jones is a short red-haired farmer. $\models$
b. Jones is red-haired.
c. Jones is a farmer.
d. Jones is not a tall red-haired farmer.

Note that it does not on the intended readings of these sentences automatically follow that 'Jones is not a tall farmer'.

5.3 *An alternative approach*

All these examples so far work, but I find in general that this method is clumsy, for a number of reasons. Firstly, we have a rather cumbersome sequence of translations to carry out: from logical form to clausal form, then to applicative form and finally to combinator form. And in order

[5] Translations and proofs tested with Prover9.

to interpret the answers we get from our first order prover we need to reverse this process, particularly for cases where we are trying to answer a wh-question. To do this adequately we need to keep track of the unifying substitutions that allow the proof to go through.

Secondly, while this is an engineering rather than a theoretical problem, it is likely that on a large scale this approach would be very inefficient at the theorem proving stage: most (particularly Prolog-inspired) theorem provers rely heavily on predicate indexing for efficient search among a large set of clauses, and all our literals have the same dummy 'p' predicate. It is easy to think of other indexing schemes that would help, but they are not necessarily straightforward to add to existing systems.

Finally, notice that in the final form of the literals we may still have logical connectives. In order to capture all the inferences associated with these (we encoded a few in a flat-footed and uneconomical way a little earlier) we would have to efficiently axiomatise the inferences associated with connectives inside lambda terms. This is a little reminiscent of what would be needed to axiomatise various forms of property theory (Chierchia *et al.* (1989); Turner (1992); Fox and Lappin (2005)) and would lead to an explosion of low level axioms that carry no weight theoretically but are disastrous computationally.

It may therefore be worth exploring an alternative approach, which combines some of the features of the techniques already described. Looking at the properties of the inference examples discussed so far it seems we need to be able to do several things:

1. replace second order terms by some kind of first order constant which retains a unique link to the second order term that it replaces,
2. be able to reason using the internal structure of the second order term where it is more complex than a predicate constant,
3. ensure that this reasoning does not go beyond FOL.

One way of achieving this is to regard our second order axioms as rewriting or translation schemata which are applied to the compositionally derived logical form in order to produce one or more "compiled" first order equivalents. This has some features of a kind of on-the-fly reification of the type discussed earlier but one which does not require pre-computation.

This leads operationally to a picture like the following:

Partly second order logical form ⇒
Second order rewriting schemata ⇒
Expanded set of first order LFs ⇒
FOL Theorem prover

We reinterpret our existing axioms as rewriting schemata: we match them to an input logical form using higher order matching (which as already remarked, is decidable), and then if necessary beta-reduce the results. We also need a "reification" function: a kind of hash function guaranteed to give a unique first order constant for each different second order argument we give it. To illustrate, we take one of our earlier axioms (there will be one for each relevant adjective of this semantic type), which says that if you are old for a P, then you are a P:

(80) $\forall xP.\ old(x,P) \rightarrow P(x)$

We reconstrue this as a rewriting rule:

(81) Adj(old): $old(x,P) \Rightarrow old(x,hash(P)) \wedge P(x)$

In order for this to give us the results we want, we have to define 'hash' as a function which produces a unique symbol of type *e* for its argument (i.e. the same argument gives the same symbol guaranteed to be unique to that argument). More on this in a moment, but first, to illustrate:

(82) Harvard is an old American university:
$old(harvard, \lambda y.\ american(y) \wedge university(y)) \Rightarrow$ (via A)
$old(harvard,*AU*) \wedge [\lambda y.\ american(y) \wedge university(y)](harvard) \Rightarrow_{beta}$
$old(harvard,*AU*) \wedge american(harvard) \wedge university(harvard)$

'*AU*' is of course the constant produced by 'hash(λy. american(y) $\wedge$ university(y))'. We can think of such constants as denoting a first-order proxy for the property described by the second order argument to 'hash', reminiscent of the output of nominalisation operators in property theory.

We cannot give a sound and complete definition for a function such as 'hash' exactly, because ideally we want it to give the same result for logically equivalent lambda-terms, and of course we cannot fully compute this logical equivalence. But we can approximate

by doing various preprocessing operations: (i) inside lambda terms, reducing expressions involving connectives to some kind of normal form, and (ii) imposing a lexicographic ordering on predicates inside disjunctions and conjunctions, so that, for example $\lambda x.P(x) \wedge Q(x)$ and $\lambda x.Q(x) \wedge P(x)$ will count as the same. There may be other useful heuristics, too: this is essentially the 'equivalence of logical form problem' often discussed in the sentence generation literature (Shieber (1993)).

We can now reinterpret our earlier axioms capturing the relation between, say, antonymous adjectives by treating all the variables in them as first order:

(83) $old(harvard,*AU*) \wedge \forall xy.old(x,y) \rightarrow \neg young(x,y) \models \neg young(harvard,*AU*)$

These can stay as axioms, added as background knowledge: they are not needed in the rewriting process.

We can deal with combinations of possessive and adjective inferences in the same way. Our main rewriting schema for possessives is now:

(84) Possessive: $of(x,y,P) \Rightarrow of(x,y,hash(P)) \wedge P(x)$

This interacts with Adj(old), an output of the rewriting schemata for adjectives, and so we have to recursively apply these rewritings:

(85) John's old wooden toy broke. =
$\exists x.of(x,john, \lambda y.old(y, \lambda z.wooden(z) \wedge toy(z))) \wedge broke(x)$

via Possessive:

(86) $\exists x.of(x,john,*OWT*) \wedge [\lambda y.old(x,\lambda z.wooden(z) \wedge toy(z))](x) \wedge broke(x)$
$\Rightarrow_{\beta}$
$\exists x.of(x,john,*OWT*) \wedge old(x,\lambda z.wooden(z) \wedge toy(z)) \wedge broke(x)$

via A:

(87) $\exists x.of(x,john,*OWT*) \wedge old(x,*WT*) \wedge [\lambda z.wooden(z) \wedge toy(z)](x) \wedge broke(x)$

which beta-reduces to:

(88) $\exists x.of(x,john,*OWT*) \wedge old(x,*WT*) \wedge wooden(x) \wedge toy(x) \wedge broke(x)$

With this machinery we can capture the following inferences:

(89) Harvard is an old American university. ⊨
Harvard is a university.
Harvard is American.
Harvard is an American university.

but, correctly, it does not follow that:

(90) Harvard is an old university.

Similarly, we can capture:

(91) John's old wooden toy broke. ⊨
A toy broke.
A wooden toy broke.

However, we cannot yet capture inferences to:

(92) a. John's toy broke.
b. John's wooden toy broke.

or inferences such as:

(93) Bush is a former US President. ⊨
Bush is a former President.

Earlier, we had an axiom which in effect distributed over conjunctions: $\forall xyPQ.of(x,y,\lambda z.P(z) \wedge Q(z)) \rightarrow of(x,y,P) \wedge of(x,y,Q)$ We could introduce an analogous axiom for privative adjectives like "former", at least for those for which inferences within their complement are transparent.[6] If we assume that "US president" is to be analysed as involving an implicit possessive, then our example will be analysed as follows:

(94) a. Bush is a former US president.
b. former(Bush,λx.of(x,US,president))

In order to capture the inference we will need an axiom like:

(95) $\forall xyP.p\text{-}adjective(x,\lambda y.of(x,y,P)) \rightarrow p\text{-}adjective(x,P)$

[6] I am at a loss to provide a characterisation of the property that will distinguish apparently valid privative inferences – like the Bush one – from the invalid "former fussy eater" to "former eater" example discussed earlier.

Noting that "US president" entails "president", and that $\lambda z.P(z) \wedge Q(z)$ entails both P and Q we might be tempted to propose more general axioms like:

(96) a. $\forall xypq.\ of(x,y,p) \wedge entails(p,q) \rightarrow of(x,y,q)$

b. $\forall xPQ.\ p\text{-}adjective(x,P) \wedge entails(P,Q) \rightarrow p\text{-}adjective(x,Q)$

However, these axioms are much *too* general. Suppose it is true that all footballers are also gamblers. Then "Smith is a former footballer" would wrongly entail that "Smith is a former gambler". Likewise, suppose it is true that all members of parliament are lawyers. Then "Smith is my member of parliament" would wrongly entail "Smith is my lawyer".[7] What we need to do is restrict the type of entailment considered to entailments valid simply on the basis of the logical forms of the second order predicates we are dealing with. This will usually involve reconstructing the inferences that are implicit in relations between our hash generated constants like *OWT*, *WT*, and so on.

Again we seem to run up against an irreducible case of inference involving second order properties. But we can, in this type of case at least, take advantage of the fact that the lambda terms involved are only a few beta-reductions away from something that is first order. If these terms are predicated of a first order entity then after beta-reduction the resulting formulae will also be first order. This suggests that we might be able to take these second order properties and use them to construct something that we can evaluate with our theorem prover.

We can write the axioms we need quite literally as:

(97) a. $\forall xypq.\ of(x,y,p) \wedge hash\text{-}entails(p,q) \rightarrow of(x,y,q)$

b. $\forall xpq.\ p\text{-}adjective(x,p) \wedge hash\text{-}entails(p,q) \rightarrow p\text{-}adjective(x,q)$

assuming that we are applying this to the output of our rewriting schemata, so that "p" and "q" are first order variables that will range over the constants generated by the "hash" function. We will treat the predicate 'hash-entails' as interpreted partly by a 'procedural attachment' (an old idea, but one recently used in natural language inference by Waldinger and Shrager (2008)) in our base theorem prover. The procedurally attached predicate will be evaluated by calling a

[7] Thanks to a referee for suggesting such examples.

separate instantiation of that theorem prover, in which any general background knowledge axioms are available, but none of the linguistically derived information related to the top level inference in which we are currently engaged.

In order to operationalise the "hash-entails" predicate we also need some housekeeping. We need the "hash" function to record the connection between the second order term it takes as input and the first order constant it gives as output. We will assume that this is achieved via a predicate recording the inputs and outputs to the hash function during the application of the various schemata described above, e.g.

(98) hashOutput(*AU*, λx.American(x) $\wedge$ university(x))

It is necessary to do this recursively to include any other hash-generated constants, for example:

(99) a. hashOutput(*OWT*, λA.old(A,*WT*) $\wedge$ wooden(A) $\wedge$ toy(A))
b. hashOutput(*WT*, λA.wooden(A) $\wedge$ toy(A))

We will also have need of a default case for those sentences in which the second order argument of 'of' or adjectives is not itself complex, as for example:

(100) hashOutput(toy_e, toy_{et})

We can now define the 'hash-entails' predicate as follows:

(101) $\forall p_e q_e R_{et} S_{et}$. hash-entails(p,q) $\Longleftrightarrow$
hashOutput(p,R) $\wedge$ hashOutput(q,S) $\wedge$ prove($\neg$($\exists$x.R(x) $\wedge$ $\neg$(S(x))))

where 'prove' represents a call to a separate instance of our theorem prover as described above. The assumption is that 'R(x)' and 'S(x)' etc. represent a full beta reduction of 'R' and 'S' applied to 'x', so that the formula to be proved ends up as strictly first order.

Now the inference we want will go through, with logical forms as shown:

(102) a. John's old wooden toy disappeared.
b. $\exists$x.of(x,john,*OWT*) $\wedge$ old(x,*WT*) $\wedge$ wooden(x) $\wedge$ toy(x) $\wedge$ disappear(x)

(103) a. Did John's toy disappear?
b. $\exists$x.of(x,john,toy) $\wedge$ toy(x) $\wedge$ disappear(x)

The relevant instantion of the axiom involving 'hash-entails' will be:

(104) of(x,John,*OWT*) ∧ hash-entails(*OWT*,toy) → of(x,John,toy)

The definition of "hash-entails" will give us:

(105) hashOutput(*OWT*,λx.old(x,*WT*) ∧ wooden(x) ∧ toy(x)) ∧
hashOutput(toy,toy) ∧
prove(¬(∃y. (old(y, *WT*) ∧ wooden(y) ∧ toy(y)) ∧ (¬toy(y))))

and the inference that calling the procedural predicate 'hash-entails' checks via 'prove' will be:

(106) ¬(∃y.(old(y,*WT*) ∧ wooden(y) ∧ toy(y)) ∧ ¬(toy(y)))

which is clearly almost trivially valid.

There is a minor wrinkle in applying the axiom concerning privative adjectives. Recall that we accounted for the invalidity of the inference from "Bush is a former US president" to "Bush is a president" by not allowing the axiom ∀xP.adj(x,P) → P(x) to apply to such adjectives. In our rewriting framework this means that we do not rewrite adj(x,P) as adj(x,hash(P)) ∧ P(x). Since "hash-entails" is a relation between first-order entities, there will be nothing for it to work with. The solution is to add a rewrite specific to this class of adjectives to yield former(Bush, *USP*). We will also need to arrange for "hashOutput" to recursively apply even where the usual rewriting has not taken place, to give:

(107) a. hashOutput(*USP*,λx.of(x,US,president_{et}))
b. hashOutput(president_{e},president_{et})

This approach has been fully implemented using a unification grammar to produce the logical forms, and a combination of two resolution theorem provers to carry out the inferences, with one being called during the evaluation of the "entails" predicate. Appendix gives a FraCaS style corpus of the natural language inferences described in this paper, all of which are successfully handled by this system.

6 CONCLUSIONS

There are some simple but quite central linguistic constructs that seem to need second order inference. It may be possible to reduce the nec-

essary inferences to those capturable in first order logic via some standard variant of reification, but the apparent requirement for a potentially infinite number of types of new first order individuals caused by recursive adjective modification seems a barrier to this. An alternative approach using translation to FOL via combinators may work, but is a little clumsy and may not generalise fully.

A better approach seems to be to pre-process the second order logical forms using second (or perhaps higher) order matching, rewriting in a forward-chaining manner to produce first-order logical forms in which second order arguments are represented by first order constants: a different type of reification, in some sense. The inferential content of these particular originally second order terms can be recaptured via a subsidiary set of first order inferences using a procedurally attached predicate which calls a separate instance of a theorem prover, after suitable beta-reductions produce first order forms.

It is an interesting question as to what extent this strategy, or variants of it, can be used to handle other types of second or perhaps higher order inference. Extensions to cover various forms of the comparative construction seem straightforward. It remains to be seen whether other second order inference phenomena such as intensional verbs may also yield to this approach.

7 ACKNOWLEDGEMENTS

This paper has been a long time in gestation. Talks based on parts of it have been given to meetings of the MOLTO project and to the Controlled Natural Language conference in Zurich, 2012; to the Computational Linguistics seminars at Kings College London and at Oxford University, 2013; and to Nuance Research Labs in Sunnyvale, CA, in 2013. I am grateful for comments received on all these occasions (particularly from Emmon Bach in Oxford), for some suggestions from Ash Asudeh, and also to Johan Bos for useful discussion of these and similar topics over many years. Four referees and two editors made many helpful criticisms and suggestions for which I am also grateful.

APPENDIX

Phenomenon	Expected answer
Intersective adj:	
Jones is a Welsh musician.	
Is Jones Welsh?	Yes
Is Jones a musician?	Yes
All musicians are teachers.	
Is Jones a teacher?	Yes
Is Jones a Welsh teacher?	Yes
Gradable adj:	
Mickey is a large mouse.	
Is Mickey a mouse?	Yes
Is Mickey large?	No proof found
All mice are animals.	
Is Mickey an animal?	Yes
Is Mickey a large animal?	No proof found
All mice are small animals.	
Is Mickey a small animal?	Yes
Mickey isn't a large animal?	Yes
Mickey isn't a small mouse?	No proof found
Privative type 1:	
Jones is a former Welsh minister.	
Is Jones a minister?	No proof found
Is Jones a Welsh minister?	No proof found
Is Jones a former minister?	Yes
Jones isn't a Welsh minister?	No proof found
Privative type 2:	
Jones owns a fake diamond.	
Does Jones own a diamond?	No proof found
Zirconia is a fake diamond.	
Zirconia isn't a diamond?	Yes
Interaction with antonyms:	
John is a tall man.	
John isn't a short man?	Yes
Bill isn't a short man.	
Is Bill a tall man?	No proof found

Recursive adj modification:

Harvard is an old American university.	
Harvard is a university?	Yes
Harvard is an American university?	Yes
Harvard is an old university?	No proof found

Possessives:

John's brother is Bill.	
Bill is a doctor.	
Is Bill John's doctor?	No proof found
Smith is Bill's plumber.	
Is Smith a plumber?	Yes
Smith is a decorator.	
Is Smith Bill's decorator?	No proof found
Smith's essay's title intrigued Jones.	
An essay's title intrigued Jones?	Yes
A title intrigued Jones?	Yes
An essay intrigued Jones?	No proof found
Smith intrigued Jones?	No proof found

Combination of adj and possessive:

John's old wooden toy broke.	
Did John's toy break?	Yes
Did John's wooden toy break?	Yes
Did John's old toy break?	No proof found
Did an old wooden toy break?	Yes
Did an old toy break?	No proof found
Did a wooden toy break?	Yes
Did a toy break?	Yes

REFERENCES

Marilisa AMOIA and Claire GARDENT (2007), A First Order Semantic Approach to Adjectival Inference, in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 185–192, Association for Computational Linguistics, Prague, http://www.aclweb.org/anthology/W07-1430.

Franz BAADER, Diego CALVANESE, Deborah L. MCGUINNESS, Daniele NARDI, and Peter F. PATEL-SCHNEIDER, editors (2003), *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press, New York, NY, USA, ISBN 0-521-78176-0.

Johan BOS (2009), Computing Genitive Superlatives, in *Proceedings of the Eighth International Conference on Computational Semantics*, IWCS-8 '09, pp. 18–32, Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-90-74029-34-6, http://dl.acm.org/citation.cfm?id=1693756.1693763.

Johan BOS, Stephen CLARK, Mark STEEDMAN, James R. CURRAN, and Julia HOCKENMAIER (2004), Wide-Coverage Semantic Representations from a CCG Parser, in *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, pp. 1240–1246, Geneva, Switzerland.

Stergios CHATZIKYRIAKIDIS and Zhaohui LUO (2014), Natural Language Inference in Coq, *Journal of Logic, Language and Information*, 23(4):441–480, ISSN 0925-8531, doi:10.1007/s10849-014-9208-x, http://dx.doi.org/10.1007/s10849-014-9208-x.

Gennaro CHIERCHIA, Barbara H. PARTEE, and Raymond TURNER (1989), Introduction, in Gennaro CHIERCHIA, Barbara H. PARTEE, and Raymond TURNER, editors, *Properties, Types and Meaning. Volume I: Foundational Issues*, pp. 1–16, Kluwer, Dordrecht.

Robin COOPER, Dick CROUCH, Jan VAN EIJCK, Chris FOX, Josef VAN GENABITH, Jan JASPARS, Hans KAMP, David MILWARD, Manfred PINKAL, Massimo POESIO, and Steve PULMAN (1996), *Using the Framework*, LRE 62-051, The FraCaS Consortium, ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz.

Donald DAVIDSON (1967), The Logical form of Action Sentences, in Nicholas RESCHER, editor, *The Logic of Decision and Action*, pp. 81–95, University of Pittburgh Press: Pittsburgh.

Jos DE BRUIN and Remko SCHA (1988), The Interpretation of Relational Nouns, in *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pp. 25–32, Association for Computational Linguistics, Buffalo, New York, USA, doi:10.3115/982023.982027, http://www.aclweb.org/anthology/P88-1004.

Guillermo DEL PINAL (2015), Dual Content Semantics, Privative Adjectives, and Dynamic Compositionality, *Semantics and Pragmatics*, 8(Article 7):1–53.

Chris Fox and Shalom Lappin (2005), *Foundations of Intensional Semantics*, Blackwell.

Jerry R. Hobbs (1985), Ontological Promiscuity, in *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pp. 60–69, Association for Computational Linguistics, Chicago, Illinois, USA, doi:10.3115/981210.981218, http://www.aclweb.org/anthology/P85-1008.

Gerard Huet (1975), A Unification Algorithm for Typed λ-Calculus, *Theoretical Computer Science*, 1:27–57.

Joe Hurd (2002), An LCF-Style Interface between HOL and First-Order Logic, in Andrei Voronkov, editor, *Automated Deduction - CADE-18, 18th International Conference on Automated Deduction, Copenhagen, Denmark, July 27-30, 2002, Proceedings*, volume 2392 of *Lecture Notes in Computer Science*, pp. 134–138, Springer, ISBN 3-540-43931-5.

Johannes A. W. Kamp (1975), Two Theories about Adjectives, in Edward L. Keenan, editor, *Formal Semantics of Natural Language*, pp. 123–155, Cambridge University Press, Cambridge.

Daniel Lassiter and Noah D. Goodman (2017), Adjectival Vagueness in a Bayesian Model of Interpretation, *Synthese*, 194(10):3801–3836, doi:10.1007/s11229-015-0786-1, https://doi.org/10.1007/s11229-015-0786-1.

Bill MacCartney and Christopher Manning (2008), Modeling Semantic Containment and Exclusion in Natural Language Inference, in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 521–528, Coling 2008 Organizing Committee, http://aclweb.org/anthology/C08-1066.

William McCune (2005–2010), Prover9 and Mace4, http://www.cs.unm.edu/~mccune/prover9/.

Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki (2015), Higher-order Logical Inference with Compositional Semantics, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2055–2061, Association for Computational Linguistics, Lisbon, Portugal, https://aclweb.org/anthology/D15-1244.

Marcin Morzycki (2014), Modification, https://www.msu.edu/~morzycki/work/papers/modification_book.pdf, accessed Jan 3 2015.

Barbara H. Partee (2007), Compositionality and Coercion in Semantics: The Dynamics of Adjective Meaning, in Gerlof Bouma, Irene Krämer, and Joost Zwarts, editors, *Cognitive foundations of interpretation*, p. 145–161, University of Chicago Press.

Barbara H. PARTEE and Vladimir BORSCHEV (2003), Genitives, Relational Nouns, and Argument-modifier Ambiguity, in Ewald LANG, Claudia MAIENBORN, and Cathrine FABRICIUS-HANSEN, editors, *Modifying Adjuncts*, Interface Explorations, pp. 67–112, Mouton de Gruyter, Berlin.

Stanley PETERS and Dag WESTERSTÅHL (2006), *Quantifiers in Language and Logic*, Clarendon Press, Oxford.

Stanley PETERS and Dag WESTERSTÅHL (2013), The Semantics of Possessives, *Language*, 89(4):713–759.

Jessica RETT (2014), *The Semantics of Evaluativity*, Oxford Studies in Theoretical Linguistics, Oxford University Press.

Stuart M. SHIEBER (1993), The Problem of Logical-form Equivalence, *Computational Linguistics*, 19(1):179–190, ISSN 0891-2017, `http://dl.acm.org/citation.cfm?id=972450.972460`.

Mark STEEDMAN (2012), *Taking Scope*, MIT Press.

Colin STIRLING (2010), Introduction to Decidability of Higher-Order Matching, in Luke ONG, editor, *Foundations of Software Science and Computational Structures*, volume 6014 of *Lecture Notes in Computer Science*, pp. 1–1, Springer Berlin Heidelberg, ISBN 978-3-642-12031-2, doi:10.1007/978-3-642-12032-9_1, `http://dx.doi.org/10.1007/978-3-642-12032-9_1`.

Zoltán Gendler SZABÓ (2017), Compositionality, in Edward N. ZALTA, editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, summer 2017 edition.

Raymond TURNER (1992), Properties, Propositions and Semantic Theory, in Mike ROSNER and Rod JOHNSON, editors, *Computational Linguistics and Formal Semantics*, pp. 159–180, Cambridge University Press, Cambridge.

Richard WALDINGER and Jeff SHRAGER (2008), Answering Science Questions: Deduction with Answer Extraction and Procedural Attachment, *AAAI Spring Symposium: Semantic Scientific Knowledge Integration*.

German particle verbs: compositionality at the syntax-semantics interface

Stefan Bott and Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Germany

ABSTRACT

Particle verbs represent a type of multi-word expression composed of a base verb and a particle. The meaning of the particle verb is often, but not always, derived from the meaning of the base verb, sometimes in quite complex ways. In this work, we computationally assess the levels of German particle verb compositionality by applying distributional semantic models. Furthermore, we investigate properties of German particle verbs at the syntax-semantics interface that influence their degrees of compositionality: (i) regularity in semantic particle verb derivation and (ii) transfer of syntactic subcategorization from base verbs to particle verbs. Our distributional models show that both superficial window co-occurrence models as well as theoretically well-founded syntactic models are sensitive to subcategorization frame transfer and can be used to predict degrees of particle verb compositionality, with window models performing better even though they are conceptually and computationally simpler.

Keywords: particle verbs, multi-word expressions, compositionality, distributional semantics

1 INTRODUCTION

Particle verbs (PVs), such as the German *auf|essen* (*to eat up*) and the English *to blow up*, represent a type of multi-word expression (MWE) composed of a base verb (BV) and a particle. While particle verbs exist in many languages, German PVs are particularly frequent and form a

highly productive paradigm which often produces neologisms and is subject to creative language use in puns and word plays.

German PVs, similarly to other MWEs, exhibit a varying degree of compositionality, as illustrated in examples (1) vs. (2). The meaning of the highly compositional PV *nach*|*drucken* (*to reprint*) is closely related to its BV *drucken* (*to print*), while the PV *nach*|*geben* (*to give in*) has little meaning in common with the BV *geben* (*to give*).

(1) *Der Verlag* *DRUCKTE* *das Buch* *NACH*.
the publisher PRINTED the book PRT_{nach}
'The publisher reprinted the book.'

(2) *Peter* *GAB* *ihrer Bitte* *NACH*.
Peter GAVE her request PRT_{nach}
'Peter gave in to her request.'

From a computational point of view, addressing the compositionality of PVs (and multi-word expressions in general) is a crucial ingredient for lexicography and Natural Language Processing (NLP) applications, in order to know whether the expression should be treated as a whole or as the sum of its constituents, and what the expression means. For example, studies such as Cholakov and Kordoni (2014), Weller *et al.* (2014) and Cap *et al.* (2015) have integrated the prediction of multi-word compositionality into statistical machine translation.

Assessing PV compositionality requires one to assess the semantic contributions of both the BV and the verb particle (Lechler and Roßdeutscher 2009; Haselbach 2011; Kliche 2011; Springorum 2011). This is obvious in highly compositional cases as in example (1): the meaning of *nach*|*drucken* (*to reprint*) is a straightforward composition of the meanings of *nach* (*again*) and *drucken* (*to print*).[1] Non-compositional cases such as *nach*|*geben* in example (2) behave differently: they are not semantically transparent with respect to the meaning of the BV, and the meaning contributed by the particle *nach* is not straightforward.

Compositionality is not a binary property of PVs, however. The levels of compositionality are distributed over a continuous scale,

[1] An evident problem is that the particle *nach* here means more than simply *again*: it implies that an additional copy is created. In addition, *nach*, like most particles, is semantically ambiguous. These issues will be addressed below.

where examples (1) and (2) refer to two extremes of the continuum, rather than prototypical cases. In contrast, *ab*|*segnen* in (3) represents an example which is judged as semi-compositional by human raters, meaning *to approve* rather than *to bless*.

(3) *Der Chef* SEGNETE *die Pläne* AB.
the boss BLESSED the plans PRT_{ab}
'The boss approved the plans.'

In this article, we investigate the factors that influence the prediction of PV compositionality from a corpus-linguistic perspective. We start with a series of hypotheses that are then investigated by a series of experiments. First, we argue that PVs can be grouped into semantically coherent classes that share the same semantic derivation when BVs from the same class are combined with a certain particle type. This combination typically selects a specific sense of the particle. Second, we address the prediction of compositionality by applying distributional semantic methods. After verifying a novel approach to model syntactic subcategorization changes, we compare window-based models with models that integrate syntactic transfer. Our main contributions are at the interface between a theoretical study of PV compositionality and the computational use of distributional semantic methods, to identify a theoretically reliable and computationally useful framework.

2 MOTIVATION AND HYPOTHESES

In this section we describe the theoretical foundations of our assumptions and analyses. We first discuss in more detail the notions of PV compositionality (Section 2.1), semantic derivation (Section 2.2), and syntactic transfer (Section 2.3). Section 2.4 then describes our distributional semantic approach, and Section 2.5 defines our hypotheses.

2.1 *Particle verb compositionality*

We illustrated above that compositionality is a scalar property: Apart from highly compositional PVs such as *nach*|*drucken*, PVs such as *ab*|*segnen* are not fully transparent with respect to their BVs, but still integrate meaning components attributed by the particle and the BV.

We refer to PVs that are semantically related to their BVs (in contrast to non-compositional PVs, which are semantically unrelated to their BVs) as *semantically derived PVs*.

Semantic derivation takes place not only for highly frequent PVs but also for infrequent or domain-specific PVs as well as neologisms. For example, while *nach*|*schneiden* in (4a) is a common verb in everyday language, *nach*|*sägen* in (4b) is more restricted to a specific domain and much less frequent; *nach*|*töten* in (4c) is a neologism.[2] The meanings of all three PVs in (4) are semantically derived from the meanings of the respective BVs, and the meaning contribution of the particle is productive and regular: All of the *nach*-PVs in (4) have a common semantic component which implies some kind of *correction to a previous action of BV by performing BV again*.

(4) a. *Der Friseur SCHNITT ihr die Haare NACH.*
the hairdresser CUT her the hair PRT_{nach}
'The hairdresser trimmed her hair.'

b. *Einfach mit der richtigen Größe NACH|SÄGEN ist nicht.*
simply with the right size PRT_{nach}|SAW is not
'You cannot simply resaw it with the right size.'

c. *Das Reh war noch nicht tot und wurde NACH|GETÖTET.*
the deer was yet not dead and was PRT_{nach}|KILLED
'The deer was not dead yet and had to be finished off.'

The same BVs from (4) can also combine with other particles, such as *an*, and undergo a different but also regular semantic derivation, as illustrated in (5). Here, all of the *an*-PVs have a common semantic component that refers to a partitive meaning, *to start a first bit of BV*.

(5) a. *Du musst das Messer abwaschen, bevor du das nächste Stück Torte AN|SCHNEIDEST.*
you must the knife clean before you the next piece cake PRT_{an}|CUT
'You have to clean the knife before you start cutting the next piece of the cake.'

[2] Examples with PV neologisms are taken from a sentence generation experiment by Springorum *et al.* (2013a), where the experiment participants generated sentences for existing and non-existing PVs.

b. *Max und Moritz SÄGEN die Brücke AN.*
Max and Moritz SAW the bridge PRT_{an}
'Max and Moritz start sawing the bridge.'

c. *Bring ihn nicht gleich um. Du solltest ihn erst AN|TÖTEN.*
bring him not already PRT_{um} you shall him first PRT_{an}|KILL
'Don't kill him right away. You should start killing him first.'

Often, similar semantic derivations apply to semantically similar BVs, such as *schneiden* and *sägen* in examples (4) and (5), which both refer to a cutting event. In these cases, we find regular semantic shifts, where combining semantically similar BVs with specific particle types results in semantically similar PVs (Springorum *et al.* 2013b; Köper and Schulte im Walde 2018). We refer to these regular semantic shifts as *semantic transfer patterns*.

(6) *Semantic Transfer Pattern*
Taking a BV from semantic group α and a particle β with meaning μ, we will derive a PV from semantic group δ.

Note that it is not the particle type that is responsible for the meaning shift, but a particular sense μ of the particle type. For example, the particle *nach* is ambiguous and does not only mean *again* (roughly corresponding to the English prefix *re*, cf. Haselbach 2011). Accordingly, the meaning of a PV may be ambiguous along the lines of the senses of the particle.

In contrast to semantically derived PVs, we refer to completely non-compositional PVs as fully lexicalized, such as *nach|geben* in (2) and *um|bringen* (*to kill*, while the BV *bringen* means *to bring*). Without diachronic considerations, the meanings of these PVs cannot directly be inferred from the meanings of their verbal bases *geben* and *bringen* and the meanings of the verb particle types *um* and *nach*.

Treating each PV as an independent lexical entry would require a large number of unrelated lexical entries and thus disregard generalizations about the semantic classes of PVs and the meaning contributions of the verb particles. Further on, a pure lexical listing approach does not explain the productivity of the PV paradigm regarding

neologisms, whose meanings are derived from regular semantic transfer patterns. The semantic pattern approach is therefore appealing, since it reduces idiosyncracy in the lexicon, and accounts for the productivity of German PVs and the ease of native speakers to produce and interpret PV neologisms.

2.2 *Semantic derivation and the meanings of particles*

What is the meaning of verb particles? Some particle senses are parallel to homophonic prepositions or adverbs (Stiebels 1996). But it is not clear if such a treatment can be extended to all particles and particle meanings. It is thus difficult to assign particles a lexical entry rather than taking whole PVs into account (Lechler and Roßdeutscher 2009; Kliche 2011; Springorum 2011).

For a more comprehensive example, consider the particle *an*. PVs with *an* can express, among other things, a direction of an action, a fixation, a manner of communication, and a partitive event, as exemplified in (7a–d) (Springorum 2011; Bott and Schulte im Walde 2014a). The particle is highly ambiguous, and its meanings are sometimes difficult to capture, but assuming (6) *Semantic Transfer Patterns* ties them closely to common underlying semantic derivations.

(7) a. *A BLICKT/SCHAUT/STARRT/STIERT B AN.*
A LOOKS/STARES/GAZES B PRT_{an}
'A looks/stares/gazes at B.'

b. *A BRÜLLT/FAUCHT/BELLT/MECKERT B AN.*
A ROARS/HISSES/BARKS/BLEATS B PRT_{an}
'A brawls/hisses/scolds at B.'

c. *A KLEBT/HEFTET/SCHRAUBT B an C AN.*
A GLUES/AFFIXES/SCREWS B at/onto C PRT_{an}
'A glues/affixes/screws B onto C.'

d. *A SCHNEIDET/BRICHT/REIßT B AN.*
A CUTS/BREAKS/TEARS B PRT_{an}
'A cuts/breaks/tears the first piece of B.'

The semantic class of the PV and individual particle meanings are also tied together by specific selectional restrictions. This is most ap-

parent in cases like (7d): the particle *an* refers to *the first bit of BV*, which is only applicable if the BV belongs to a semantic class that allows for a partitive meaning, such as *consumption, cutting,* etc. Also, it is not trivial to decide if two PVs share the same sense of a particle or not, as in (7a) vs. (7b). Does *an* only express some kind of directionality or are the two semantic transfer patterns sufficiently different to assume two particle meanings? Note that our definition of semantic derivation does not make any claim about how to discriminate between particle senses and how to establish a number of senses.

The ambiguity of particles often leads to different senses of PVs, even if the PVs are compositional with respect to the same meaning of the BV. For example, the PV *an|fahren* can have at least three meanings. It is ambiguous between *to drive into* as in (8a), *to start driving* as in (8b), and *to approach by driving* as in (8c). These particle meanings of *an* are shared among semantically similar PVs, respectively, e.g., *an|rempeln* (*to bump into*), *an|laufen* (*to start running*) and *an|steuern* (*to approach by steering*, e.g. a ship).

(8) a. *Das Auto FUHR den Fußgänger AN.*
the car DROVE the pedestrian PRT_{an}
'The car ran into the pedestrian.'

b. *Das Auto FUHR AN, als die Ampel grün wurde.*
the car DROVE PRT_{an}, when the light green turned
'The car went when the light turned green.'

c. *Der Bus FUHR die Haltestelle AN.*
the bus DROVE the stop PRT_{an}
'The bus approached the bus stop.'

We also find cases where a new non-standard meaning is enforced by the semantic interpretation of a PV. (9) is an example from an advertisement campaign for a soft drink which carries the word *Sonne* (*sun*) in its name. Here the PV *zu|gehen* (*to close*) is used, along with the PV *auf|gehen* (*to rise* and *to open*). The sun cannot *close*, but the new type of package – which is advertised here – can.

(9) *Die Sonne GEHT AUF.* Und *ZU.*
the sun GOES PRT_{auf} and PRT_{zu}
'The sun rises/opens. And closes.'

A definition of particle meaning in terms of semantic transfer patterns as expressed by (6) is compatible with all of the findings listed above, while it does not define precise lexical entries for particles and does not make claims about the number of senses per particle.

2.3 *Syntactic transfer*

So far, we have only discussed the semantic aspects of PVs, but the shifts from BVs to PVs also influence the syntactic behavior of the PVs, which in turn may provide a helpful approximation to the semantics of PVs (Levin 1993). To illustrate the syntactic aspect, consider the examples in (10). Although the PV *an|leuchten (to shine at)* is rather compositional, the means for the illumination *Lampe (lamp)* is represented by the subject of the BV in (10a) vs. a PP complement headed by mit_{dat} of the PV in (10b). PV and BV thus behave syntactically differently with respect to their argument structures and the syntactic functions of identical semantic roles.

(10) a. *Die Lampe* LEUCHTET.
the lamp shines
'The lamp shines.'

b. *Peter* LEUCHTET *das Bild mit der Lampe* AN.
Peter SHINES the picture with the lamp PRT_{an}
'Peter illuminates the picture with the lamp.'

In addition to changes in the predominant syntactic functions for semantic arguments when comparing PVs to their BVs, we also find *extension* and *incorporation* of syntactic complements, as illustrated by (11) and (12), respectively. The BV *bellen* (*to bark*) in (11) is intransitive, while the corresponding PV *an|bellen* (*to bark at*) is transitive and takes an additional accusative object to express the entity being barked at. This is a case of argument extension within PV subcategorization with respect to its BV. The PV *an|schrauben* (*to screw on*) in (12) shows argument incorporation: it rarely selects an argument to express the location onto which something is screwed, while its BV *schrauben* (*to screw*) adds a complement (here: a PP) to express the direction.

(11) a. *Der Hund* BELLT.
the dog_{nom} BARKS
'The dog barks.'

b. *Der Hund* *BELLT* *den Postboten* *AN*.
the dog_{nom} BARKS the postman_{acc} PRT_{an}
'The dog barks at the postman.'

(12) a. *Der Mechaniker* *SCHRAUBT* *die* *Abdeckung* *auf* *die* *Öffnung.*
the mechanic_{nom} screws the cover on the opening_{acc}
'The mechanic screws the cover on the opening.'

b. *Der Mechaniker* *SCHRAUBT* *die* *Abdeckung* *AN*.
the mechanic_{nom} SCREWS the cover PRT_{an}
'The mechanic fixes the cover.'

Usually, groups of verbs which are similar in meaning also have similar subcategorization frames and selectional preferences (Schulte im Walde 2000; Merlo and Stevenson 2001; Korhonen *et al.* 2003; Schulte im Walde 2006; Joanis *et al.* 2008). But in (10)–(12) we can observe that this is not necessarily the case for pairs of PVs and their BVs, even if the meaning of the PV is highly transparent.

The problem illustrated here is what we call the *syntactic transfer problem*: the subcategorization frame of the BV must be mapped onto the subcategorization frame of the PV, and the semantic arguments are not necessarily realized as the same syntactic complements by the two verbs. Note that such syntactic transfer patterns tend to be quite stable within groups of PVs with the same semantic shift (Aldinger 2004; Bott and Schulte im Walde 2014c).

One way to computationally address the syntactic transfer problem is by measuring the overlap between all complement slot combinations of any given PV–BV pair and to identify the best correspondences between the slots. We suggest distributional semantic models to support us in the assessment of PV compositionality, while paying attention to syntactic PV–BV transfer: if the PV is non-compositional, we expect a large distributional distance between the correspondences of PV–BV subcategorization slots. For example, in (13b) the PV *an*|*drehen* (*to palm off sth. on so.*) is opaque with respect to the BV *drehen* (*to turn*). The typical patients of *turning* (*drehen*) events may be *knobs*, *wheels* and *heads*, cf. (13a), which are different from the typical patients of a *selling* event as in *an*|*drehen*. We thus ex-

pect to find very different words as typical fillers of the direct object slot of the two verbs, signalling that the two slots do not express the same type of semantic argument, and that the PV is thus non-compositional.

(13) a. *Eulen können ihren Kopf nach hinten* DREHEN.
owls can their head$_{acc}$ to the back TURN
'Owls can turn their heads around backward.'

b. *Der Verkäufer hat ihm das Auto* AN|GEDREHT.
the seller has him the car$_{acc}$ PRT$_{an}$|TURNED
'The salesman has palmed the car off on him.'

The strength of the syntactic transfer will be taken as a proxy for semantic classes and compositionality. We hypothesize that the higher the distributional associative strength between the slots within a syntactic transfer pattern, the stronger the PV compositionality. We further hypothesize that the semantic transfer patterns expressed by (6) are paralleled by regular syntactic transfer patterns.

2.4 *Distributional information*

In order to test our assumptions against empirical data we use distributional semantic models. According to the distributional hypothesis, the meaning of a word is characterized by the distribution of its contexts (Harris 1954; Firth 1957). Intuitively, this corresponds to the idea that we expect to find a word such as *driver* in the context of the word *car*, and the word *captain* in the context of the word *ship*.

One way of defining the concept of *context* is a vector in a high-dimensional space, where each dimension represents an aspect of contextual distribution, such as context words (Sahlgren 2006; Turney and Pantel 2010). Each target word is represented by a vector, and each vector dimension is determined by the co-occurrence strength with context words. For example, if *bone* occurs c times in the local context of *dog*, the dimension *bone* in the vector of *dog* will be c. If each vector dimension refers to a context word, the unreduced vector space has as many dimensions as there are word types in the corpus.

It is possible to reduce the dimensionality and thus abstract over individual lexical items by applying dimensionality reduction techniques, such as Random Indexing (Sahlgren 2005), Singular Value

Decomposition (Landauer and Dumais 1997) and Latent Dirichlet Allocation (Blei *et al.* 2003). It is also possible to use more complex units of context than simple words as vector dimensions, e.g., by relying on subcategorization functions (Padó and Lapata 2007), where verbs can, for example, be characterized by the kinds of subjects or objects they typically take. An obvious example is that we expect to find *dog* as a typical subject of the verb *to bark* and *cat* as a typical subject of *to meow*. The distributional similarity/distance between two lexical items can be measured as the geometrical distance between their vectors, e.g. by computing the cosine of the angles of said vectors.

While distributional methods cannot provide clear-cut lexical definitions, they are convenient and successful proxies for comparing words semantically: words which are similar in meaning have a strong tendency to appear in similar contexts. Applied to the problem of PV compositionality, we can expect that distributional closeness of PVs and BVs signals high compositionality. For our experiments, we use the following configurations of context representations:

- *Windows* of surrounding lemmatized words: we use *n* words to the left and to the right of each target word, where *n* is a variable. Vector components represent words from the context, and the extension in each dimension represents frequency or *local mutual information* (LMI) as association strength (Evert 2004).

- *Complement slot fillers* for syntactic subcategorization models: vectors represent subcategorization slots for each verb (either BV or PV); vector components correspond to slot filler words or abstractions of slot fillers (such as latent dimensions).

- *Subcategorization frames*: dimensions represent subcategorization frames for each PV–BV pair. Each vector component corresponds to the observed frequency of a subcategorization frame. The distance between different PV–BV pairs can be used as a criterion for grouping together verb pairs with similar patterns.

From a practical point of view, the window approach has an advantage over the syntactic approach because it can use much more evidence mass: it is not restricted to verb arguments and can thus use all words in local contexts. From a theoretical point of view, however, the win-

dow approach does not integrate the linguistic generalizations we discussed above: regularity of semantic shifts and instances of syntactic transfer.

2.5 *Hypotheses*

The goal of this article is to empirically test hypotheses H1–H3 which we have derived on a theoretical basis:

H1 *Semantic Transfer:* For PVs that are not fully lexicalized there are groups of BVs which undergo the same semantic derivation when they combine with the same particle type, cf. Sections 2.1 and 2.2.

H2 *Syntactic Transfer:* The semantic transfer patterns are paralleled by syntactic transfer patterns, cf. Section 2.3.

H3 *Distributional Transfer:* The degree of PV compositionality can be assessed by comparing distributional PV and BV contexts at the syntax-semantics interface, cf. Section 2.4.

Following an overview of related previous work on particle verbs in Section 3, Section 4 will define and conduct three experiments according to our three hypotheses.

3 PREVIOUS APPROACHES TO PARTICLE VERBS

German PVs have been studied extensively from a theoretical point of view (Stiebels and Wunderlich 1994; Stiebels 1996; Lüdeling 2001; Dehé *et al.* 2002; Müller 2002, 2003; McIntyre 2007).[3] Lüdeling (2001) investigated whether PVs are morphological objects or phrasal constructions and how they can be distinguished from secondary predicate constructions or adverbial constructions. She revealed a series of theoretical problems and analyzed PVs as lexicalized phrasal constructions, considering separability the strongest argument for this analysis. Olsen (1997) studied German PVs at the morpho-syntactic interface and analyzed cases in which an explicit argument of a BV becomes implicit in the formation of a PV. Müller (2002, 2003), in turn, argued for an analysis of PVs as verbal complexes at the morpho-syntactic interface, and provided lexical interpretations. Under his view, PVs

[3] Also see a bibliography on verb particle constructions, as maintained by Nicole Dehé until 2015: `http://ling.uni-konstanz.de/pages/home/dehe/bibl/PV.html`.

are seen as both morphological and syntactic objects. For the present work, the status of PVs on the morphological vs. the syntactic level is not relevant, so we will not commit ourselves to a specific perspective in this respect.

Research addressing the semantics of verb particles has mostly focused on specific particle types, such as *auf* (Lechler and Roßdeutscher 2009), *nach* (Haselbach 2011), *ab* (Kliche 2011), and *an* (Springorum 2011). Springorum *et al.* (2012) and Rüd (2012) presented automatic classification methods for PVs with the particles *an* and *auf*, respectively. Springorum *et al.* (2013b) provided a case study of regular meaning shifts in PVs where they argue that particles have a meaning which is implicit in the semantic transfer pattern, in a similar way as we argue here.

Predicting degrees of PV compositionality from a computational perspective has been addressed previously, mainly for English. Most prominently, Baldwin *et al.* (2003) defined a word-based model of Latent Semantic Analysis for English particle verbs and their constituents, and measured the distributional similarity of the models to evaluate the resulting degrees of compositionality against various WordNet-based gold standards. McCarthy *et al.* (2003) exploited measures on syntax-based distributional descriptions as well as selectional preferences, to predict the compositionality of English particle verbs. Bannard (2005) describes a distributional approach that compared word-based co-occurrences within the British National Corpus for English particle verbs with those of the respective base verbs and particles. Cook and Stevenson (2006) addressed the compositionality and the meaning of English particle verbs by a distributional model encoding standard verb semantic features (especially subcategorization-based information) and PV-specific heuristics. A larger multifactorial study of idiomacity within a construction grammar framework (Wulff 2010) introduced a measure to compute compositionality with respect to both PV constituents.

Regarding computational approaches to German PVs, Aldinger (2004) and Schulte im Walde (2004, 2005) were the first to study them from a corpus-based perspective, with an emphasis on the subcategorization behavior and syntactic change. Aldinger (2004) investigated the regularity in syntactic subcategorization transfer. Schulte

im Walde (2005) explored salient features at the syntax-semantics interface that determined the nearest semantic neighbors of German PVs. Relying on the insights of this study, Hartmann (2008) presented preliminary experiments on modeling the subcategorization transfer of German PVs by measuring the overlap of argument heads, in order to strengthen PV–BV distributional similarity. The results of that study were not conclusive due to data sparseness. Kühner and Schulte im Walde (2010) used unsupervised clustering to determine the degree of compositionality of German PVs. They hypothesized that compositional PVs tend to occur more often in the same clusters with their corresponding BVs than opaque PVs. Their approach relied on nominal complement heads in two modes, (i) with and (ii) without explicit reference to the syntactic functions. The explicit incorporation of syntactic information (i) yielded less satisfactory results, since a given subcategorization slot for a PV complement does not necessarily correspond to the same semantic type of complement slot for the BV, thus putting the syntactic transfer problem in evidence, again.

Bott and Schulte im Walde (2014b) showed that a window-based model can predict degrees of compositionality and establish a ranking of PVs accordingly, to significantly correlate with human ratings. Within this study, we focused on the influence of various linguistic factors, such as the ambiguity and the overall frequency of the verbs and syntactically separate occurrences of verbs and particles that typically cause difficulties for the correct lemmatization of PVs.

Köper and Schulte im Walde (2017) combined similar textual distributional information with images, to improve the prediction of compositionality for German noun compounds and particle verbs. Bott and Schulte im Walde (2014c) argued that the semantic classes of PVs can be predicted by purely syntactic features. We showed that automatically derived semantic classes overlap significantly with class distinctions based on human ratings. In Bott and Schulte im Walde (2014a), we showed that a computational assessment of syntactic transfer patterns is feasible and that a computational model can predict slot correspondences. Finally, in Bott and Schulte im Walde (2015) we presented preliminary work on predicting PV compositionaly on the basis of the modeling of syntactic transfer patterns.

4 EXPERIMENTS

Up to now, we motivated our research hypotheses from a theoretical perspective. In this section, we assess our hypotheses within three computational experiments. In Section 4.1, we approximate semantic transfer and the meaning of particles by semantically clustering PVs that share semantic transfer patterns, while using syntactic features in the form of subcategorization frames. In Section 4.2, we verify that syntactic transfer can be predicted in isolation, and in Section 4.3, we compare window-based models and models integrating syntactic transfer information to determine the compositionality of PVs. The experiments presented here are based on preliminary investigations in Bott and Schulte im Walde (2014b,c,a, 2015), which we now extend and discuss in more detail and depth.

4.1 *Experiment 1: Modeling semantic transfer*

The first experiment explores semantic derivation and the meanings of particles. Based on our theoretical considerations, we expect PV–BV pairs to group such that both BVs and PVs are semantically similar, and that the relation between them (i.e. a particle meaning) is captured as a consistent semantic transfer pattern. Since we also assume that semantic derivation is reflected by syntactic transfer patterns, we aim to automatically derive semantic groups on the basis of the syntactic behavior of PV–BV pairs.

As argued above, it is difficult both to define the meanings of particles and to clearly distinguish between them. For this reason, supervised classification techniques are reasonable, as they require training and test sets which reliably reflect distinctions between particle senses. Such data sets are expensive to create, however, and it is difficult to agree on exact numbers and definitions of particle senses on theoretical grounds. For these reasons, we believe that the derivation of groups of PV–BV pairs (and different particle senses) can be addressed more efficiently by means of clustering techniques.

4.1.1 Gold standard classification

We created a gold standard of 32 PVs listed in Fleischer and Barz (2012), including 14 PVs with the particle *an* and 18 PVs with the particle *auf*. We focused on two particle types in order to have a small and controlled test bed which allows us to study the syntactic transfer

in detail. The selected verbs were considered highly compositional, in order to investigate the correspondences between subcategorization properties. The PV set contains PVs with argument slots that are typically realized through different syntactic subcategorizations, as in example (10) with *an|leuchten*. In addition, the PV set contains PVs exhibiting argument incorporation or extension. We excluded PVs which are clearly polysemous.

The full gold standard is presented in Table 1. The first part of the *semantic class* labels was taken from Fleischer and Barz (2012); we further distinguished between the classes based on the meanings of the BVs (second part of the labels), by breaking down the general classes into more detailed classes, such as verbs of *tying*, *gaze* and *sound*. The selected verbs have a clear subcategorization pattern for BVs and PVs.

In order to validate the gold standard, we assessed it with the help of six human expert raters,[4] all German native speakers with a linguistic background. The raters were not directly asked to group PVs into categories. Instead, the PVs were presented in pairs,[5] and the raters decided whether or not the pairs belonged to the same semantic category, taking semantic similarity of the PVs as the basis for their decision. For example, the PVs *an|schneiden* (*to start cutting*) and *an|ketten* (*to chain at*) were presented as a pair to be rated. In this case, the decision that they *do not belong to the same semantic class* was expected. No pre-defined categories were provided, and the raters were not asked to provide a name or description of the categories. We did *not* ask participants to take any syntactic criteria into consideration, which were the criteria we actually used for the compilation of the gold standard.

The inter-annotator agreement was substantial (Landis and Koch 1977) with Fleiss' $\kappa = 0.68$ (Fleiss 1971).[6] As a measure of agreement between raters and the previously created gold standard, we performed pair-wise calculations. For this assessment, the gold standard was transformed into PV pairs, and the value *true* was assigned if

[4] All human ratings in this article exclude the authors as raters.

[5] All possible PV combinations were generated, while keeping PVs with *an* separate from those with *auf*.

[6] One of the six raters showed low agreement with the other raters. Eliminating this rater from the calculation of agreement, we achieved an even higher inter-annotator agreement score of $\kappa = 0.76$.

Table 1: The gold standard PV–BV classes, with sub-categorization patterns

Particle	Typical frames for the BV	Typical frames for the PV	Semantic class	Verbs in class	
an	NPnom + NPacc + PP-an	NPnom + NPacc + PP-an	locative/ relational tying	an|binden an|ketten	to tie at to chain at
	NPnom + PP-zu/ in/nach/ auf	NPnom + NPacc	locative/ relational gaze	an|blicken an|gucken an|starren	to glance at to look at to stare at
	NPnom + NPacc + PP-mit	NPnom + NPacc + PP-mit	ingressive consump-tion	an|brechen an|reißen an|schneiden	start to break start to tear start to cut
	NPnom	NPnom + NPacc	locative/ relational sound	an|brüllen an|fauchen an|meckern	to roar at to hiss at to bleat at
	NPnom + NPacc + PP-an	NPnom + NPacc	locative/ relational fixation	an|heften an|kleben an|schrauben	to stick at to glue at to screw at
auf	NPnom	NPnom	locative/ blaze-bubble	auf|brodeln auf|flammen auf|lodern auf|sprudeln	to bubble up to light up to blaze up to bubble up
	NPnom + PP-zu/ in/nach/ auf	NPnom	locative/ gaze	auf|blicken auf|schauen auf|sehen	to glance up to look up to look up
	NPnom + NPacc	NPnom + NPacc	locative/ dimensional instigate	auf|hetzen auf|scheuchen	to instigate to rouse
	NPnom + NPacc + PP-auf	NPnom + NPacc	locative/ relational fixation	auf|heften auf|kleben auf|pressen	to staple on to glue on to press on
	NPnom	NPnom	ingressive sound	auf|brüllen auf|heulen auf|klingen auf|kreischen auf|schluchzen auf|stöhnen	suddenly roar suddenly howl suddenly sound suddenly scream suddenly sob suddenly moan

Table 2: Inter-annotator agreement and comparison of the gold standard and the human ratings (Fleiss' κ)

	an	auf	an + auf
Inter-annotator agreement	0.79	0.64	0.70
Average agreement between annotators and gold standard	0.73	0.74	0.73

the two verbs of a pair belonged to the same category, and *false* otherwise. κ scores were calculated for each annotator, and the average of the agreement scores was taken.

Table 2 presents the human–gold comparison, separately for *an* and *auf* and also for the gold standard as a whole. While for the particle *an* the inter-annotator agreement is higher than the agreement between raters and gold standard, the reverse is true for the particle *auf*, and on average the human agreement with the gold standard is similar to the agreement among the annotators. We conclude that our gold standard provides a valid representation of human language intuition. Most importantly, the annotators did not use syntactic criteria and still validated a gold standard whose creation was explicitly based on syntactic subcategorization frames. In other words: there is an apparent syntax-semantics relation for our selected PVs.

4.1.2 Feature selection

As basis for corpus-based features, we used a lemmatized and tagged version of the SdeWaC corpus (Faaß and Eckart 2013), a web corpus of ≈880 million words. For linguistic pre-processing, we used the MATE parser (Bohnet 2010) to extract syntactic subcategorization frames.

For each PV–BV pair, we extracted two parallel sets of features, one for the BV and one for the PV. This allowed us to model the syntactic transfer. For example, we expected that an ideal transfer from a group of transitive BVs to a group of intransitive PVs should be reflected in high values for the features *BV:transitive* and *PV:intransitive*[7] and, in turn, low values for *BV:intransitive* and *PV:transitive*.

We distinguished between two ways of selecting the feature types from the corpus: manually and automatically. For the manual feature selection, we extracted only those features from the parsed frames

[7] Note that *transitive* and *intransitive* are only convenient abbreviations for the labels *NPnom* and *NPnom + NPacc*, which are used in Table 1.

which we already used in the creation of the gold standard and which are listed in Table 1. This resulted in a small feature set of 30 features (15 features for PVs and BVs, respectively). For the automatic feature selection, we used the *n* most frequent frames in the corpus, as determined across the set of verbs in the gold standard. In order to create an artificial upper bound, we used the typical frames as defined in Table 1 as a set of idealized "lexicographic" descriptions.

Regarding the syntactic dependency representation provided by the parser, we excluded subjects and modifiers from the representation of subcategorization frames. We, however, included PP modifiers because quantitative information on PP adjuncts has proven successful next to that of PP arguments (Schulte im Walde 2006; Joanis *et al.* 2008).

The feature vectors were normalized to their unit vectors of length 1, because the frequency ratio between BVs and PVs potentially varied strongly. The vector combination for each PV–BV pair was done by simply concatenating the dimensions of the two BV and PV vectors. In this way, each subcategorization frame was represented for both the BV and the PV. For example, the vectors for the intransitive frame were represented as *BV:intransitive* and *PV:intransitive*.

4.1.3 Clustering methods

We wanted to assess and compare hard and soft clustering for our problem, so we applied the two clustering algorithms *K-means* and *Latent Semantic Classes (LSC)*. K-means is a widely used flat, hard-clustering algorithm; we used the Weka implementation (Witten and Frank 2005). LSC (Rooth 1998; Rooth *et al.* 1999) is a two-dimensional soft-clustering algorithm which learns three probability distributions: one for the clusters, and one for the output probabilities of each element and for each feature type with regard to a cluster. The latter two (elements and features) correspond to the two dimensions of the clustering. In our case the elements are the PV–BV pairs, and the features are normalized counts of the subcategorization frames.

4.1.4 Evaluation

We evaluated the clusterings in terms of *Purity* (Manning *et al.* 2008), *Rand Index* (Rand 1971) and *Adjusted Rand Index* (Hubert and Arabie 1985). Purity assesses individual clusters in terms of the ratio between

the number of elements of the majority class and the total number of elements in the data set. A perfect clustering has a Purity of 1 while the lower bound is 0. Since Purity does not capture the amount of clusters over which each target class is distributed, also non-perfect clusterings may have a Purity of 1. However, as long as the number of clusters is constant, Purity provides an intuitive means to evaluate our cluster analyses.

The Rand Index (RI) looks at pairs of elements and assesses whether they have been correctly placed in the same cluster. RI is sensitive to the number of non-empty clusters and can capture both the quality of individual clusters and the amount to which elements of target categories have been grouped together. Since RI looks at pair-wise decisions, it is also applicable to the human ratings. The Adjusted Rand Index (ARI) is a variant of RI which is corrected for chance. RI has values between 0 and 1; ARI can have negative values.

We evaluated the cluster analyses of the verbs with the particles *an* and *auf* separately and for the gold standard as a whole (*an*+*auf*). We set the number of clusters equal to the number of target gold categories: 5 clusters for both the *an*-set and the *auf*-set and 10 clusters for the whole gold standard.

For the evaluation of LSC clusters with respect to Purity, RI and ARI, we transferred each soft clustering to a hard clustering by applying a cutoff value to the output probabilities for cluster membership. We tried various cutoff levels and found that for the sets of *an* and *auf* PVs 0.1 provided a reasonable trade-off between coverage (the total number of elements retained in all clusters) and ARI. This is also the value used in Kühner and Schulte im Walde (2010) in a similar setup.

4.1.5 Results and discussion

The clustering results are presented in Table 3, with the best automatically obtained results in gray cells. The human rating scores are given in the first row and allow for a direct comparison between automatic clustering and human decisions.[8] The second row shows the upper bound represented by the manually defined feature vectors. Note that

[8] Differently to RI, Purity and ARI are not based on pair-wise decisions and thus not applicable to the human ratings.

Table 3: Results across clustering methods and feature sets

		an			auf			an + auf		
		Purity	RI	ARI	Purity	RI	ARI	Purity	RI	ARI
Human ratings			0.93			0.92			0.92	
K-means	upper: bound: idealized features	0.83	0.91	0.70	0.88	0.92	0.72	0.93	0.97	0.82
	selected features	0.67	0.82	0.29	0.75	0.87	0.52	0.46	0.88	0.32
	20 feat	0.58	0.74	0.18	0.69	0.69	0.40	0.43	0.88	0.14
	50 feat	**0.67**	**0.80**	0.20	0.75	0.83	0.38	0.43	0.90	0.19
	100 feat	**0.67**	0.79	0.18	0.75	0.83	0.40	0.49	0.90	0.21
	200 feat	0.58	0.74	0.13	**0.81**	**0.86**	0.52	0.43	0.88	0.18
LSC	selected features; cutoff: 0.1	0.63	0.78	**0.22**	0.80	0.85	**0.55**	**0.85**	**0.92**	**0.59**

this is an *artificial* upper bound and not an experimental result, even if obtained by clustering.

The third row corresponds to the evaluation results for the manually selected corpus-based features used within K-means, in comparison to the following rows concerning the results based on the automatically selected n most frequent features, with $n = \{20, 50, 100, 200\}$. The last part of the table shows the results obtained with the LSC soft clustering algorithm, when applying the cutoff of 0.1 to the cluster membership probability. Note that the Purity values are comparable to each other because the number of clusters was held constant.

The results relying on our manual features as provided by Table 1 do not get perfect scores of 1 because of lexicographic differences concerning individual entries. They are, however, highly similar to the results obtained by the human validation of the gold standard, and thus demonstrate the feasibility of our approaches. The automatic clustering results relying on corpus-based features result in lower scores, of course, but they still represent a very strong tendency to group together PV–BV pairs into semantic classes. We can achieve relatively high Purity and RI scores, thus demonstrating that our approach is generally valid.

Concerning the corpus-based features, the manually selected set seems to perform only slightly better than the automatic feature selection settings. This is surprising, since the manually selected set was "tuned" to use the most salient features for our task. So while the noise adds potentially unrelated features, it does not considerably harm the cluster analyses. There appears to be no optimal setting for n to provide the best results across all settings. It is clear from the table, however, that the lowest number of features ($n = 20$) tends to be outperformed by a larger number of features.

As a general tendency, the soft clusterings by LSC perform on a comparable level with the hard clusterings by K-means. For the joint gold standard set *an* + *auf* and a cutoff point of 0.1, LSC performs even much better than K-means. But this comes at the cost of a very low coverage: Only 20 verbs are retained in the converted clusters, while the target size is 32.

Given that (i) the automatic clustering was performed on the basis of syntactic features while the annotators in the human classification task focused on purely semantic criteria, and that (ii) the cluster analyses were rather successful, we conclude that the semantic and the syntactic perspectives led to the creation of similar classes. We therefore provided empirical evidence for both hypotheses H1 and H2.

4.2 *Experiment 2: Modeling syntactic transfer*

In Section 2, we hypothesized that syntactic transfer patterns can be detected with distributional methods. If subcategorization slots from a PV–BV pair correspond to each other and realize the same semantic argument, we expect them to be distributionally similar. This hypothesis was tested with the following experiment.

4.2.1 Automatic prediction of slot correspondences

We rely on the same gold standard as in the previous experiment (cf. Table 1). Most importantly, the dataset contains PV–BV verb pairs whose argument slots are typically realized by different syntactic subcategorizations, as described by the expected "typical frames". The differences in the typical frames for PV vs. BV groups represent the expected transfer patterns.

The aim of this experiment was to predict transfer patterns by correspondences between syntactic slots in PV and BV subcategoriza-

tion frames. Firstly, we extracted all subcategorization frames for both BVs and PVs from the parsed version of the SdeWaC corpus. We then selected the n most frequent subcategorization frames, where n was limited to 5. Each of these frames is a set of subcategorization slots of the form $\{\sigma_1, \ldots, \sigma_m\}$. If $frame_{v,i}$ refers to the set of subcat slots of the i^{th} most frequent subcategorization frame for a verb v, we then define the set $slots_{v,n}$ as follows:

(14) $$slots_{v,n} := \{\sigma_j | \sigma_j \in frame_{v,i}, 0 < i \leq n\}$$

Informally, $slots_{v,n}$ is the set of subcat slots which appear in any of the n most frequent frames of v. The simple transitive frame, for example, contains a subject slot and an accusative object slot.

We built a vector space model for all possible combinations of BV slots and PV slots for each PV–BV pair $<pv, bv>$. The dimensions of the vector were instantiated by the head nouns of the respective syntactic function. The best matching slot $\hat{\sigma}'$ of a PV for a given slot σ_i (with slot vector $\vec{\sigma}_i$) of the corresponding BV is then defined as the maximum slot cosine score:

(15) $$\hat{\sigma}' := \underset{\sigma_j | \sigma_j \in slots_{pv,n}}{\arg\max} \cos(\vec{\sigma}_i, \vec{\sigma}_j)$$

Table 4 shows the most frequent dimensions in the vectors corresponding to PP arguments headed by *an* for the verbs *heften* (*to attach*) and *an|heften* (*to attach to*). The two verbs can be used in similar contexts with similar arguments. For example, both vectors include head nouns expressing typical places to attach things to, such as a *pin board* (*Pinnwand*), a *wall* (*Wand*), and a *board* (*Brett*). Accordingly, the two vectors are similar to each other. Note that although both vectors correspond to PP slots headed by the preposition *an*, a syntactic transfer from the accusative to the dative case takes place. In addition, the example vectors demonstrate that the features are often sparse.

A variable threshold was applied to the cosine similarity, to separate corresponding from non-corresponding subcategorization slots. This is important for the detection of argument incorporation and extension. If, for example, for a given BV slot no PV slot can be found with a cosine value above the threshold, we interpret this as a case of argument incorporation. In contrast, a slot from a PV which cannot be matched to a slot of its BV is taken to signal argument extension.

Table 4: Most frequent dimensions for two sample vectors representing subcategorization slots of the verbs *heften* (*to attach*) and *an|heften* (*to attach to*)

anheften-an$_{dat}$	count	**heften-an**$_{acc}$	count
Oberfläche (surface)	3	Ferse (heel)	154
Gerichtstafel (court notice board)	3	Brust (breast)	48
Stelle (spot)	2	Revers (lapel)	43
Schluss (end)	2	Kreuz (cross)	32
Unterlage (document)	1	Wand (wall)	30
Kirchentür (church door)	1	Spur (trace)	12
Brett (board/plank/shelf)	1	Tafel (board)	11
Pinnwand (pin board)	1	Fahne (flag)	11
Körper (body)	1	Tür (door)	11
Punkt (point)	1	Pinnwand (pin board)	9
Bauchdecke (abdominal wall)	1	Kleid (dress)	6
Baum (tree)	1	Brett (board/plank)	6
Schleimhautzelle (epithelial cell)	1	Mastbaum (mast tree)	6
Himmel (heaven/sky)	1	Körper (body)	5
Spur (trace)	1	ihn (him)	5
Sphäre (sphere)	1	Kleidung (clothing)	5
Wand (wall)	1	Oberfläche (surface)	5
Hauptreaktor (main reactor)	1	Stelle (spot)	4
Engstelle (constriction)	1	Baum (tree)	4
Pflanze (plant)	1	Jacke (jacket)	4
Protein (protein)	1	Mantel (coat)	4
Unterseite (down side)	1	Teil (part)	3
Zweig (twig)	1	Krebszelle (cancer cell)	3
Geist (spirit)	1	mich (me)	3
Pin-Wand (pin board)	1	schwarz (black)	3

For initializing the BV and PV vector dimensions, we relied on the subcategorization database compiled by Scheible *et al.* (2013), which provides a convenient access to subcategorisation information in the same dependency-parsed version of the SdeWaC corpus as used in the previous experiment. Once the verb vectors were built, we used them to predict subcategorization transfer. The baseline for the predictions was obtained by a random PV–BV slot correspondence. The results will be presented in Section 4.2.3, after introducing the gold ratings.

4.2.2 Human ratings on slot correspondences

Each pair of subcategorization slots described in Section 4.2.1 was rated by human judges. The pairs were presented as

<BV-subcategorization-slot, PV-subcategorization-slot>

and in blocks corresponding to identical BV subcategorization slots, such that the raters could directly compare all PV subcategorization slots for a given BV slot. The order of the blocks was randomized.

The raters were asked to rate the pairs on their semantic correspondence. Three annotation examples were provided to guide the ratings, cf. (16). (16a) presents a negative example, as no grammatically correct sentence is possible for *durch*|*schwimmen* with a PP complement headed by *durch*. Accordingly, the sentence in (16a-iii) is ungrammatical. (16b) presents a positive example. In unclear cases, the raters were invited to produce example sentences.

(16) a. (i) <schwimmen-durch$_{acc}$, durchschwimmen-durch$_{acc}$>
(ii) *Der Hund* SCHWIMMT *durch den Fluss.*
the dog SWIMS through the river$_{acc}$
'The dog swims through the river.'
(iii) **Der Hund* DURCH|SCHWIMMT *durch den Fluss.*
the dog PRT$_{durch}$|SWIMS through the river$_{acc}$

b. (i) <schwimmen-durch$_{acc}$, durchschwimmen$_{acc}$>
(ii) identical to (16a-ii)
(iii) *Der Hund* DURCH|SCHWIMMT *den Fluss.*
the dog PRT$_{durch}$|SWIMS the river$_{acc}$
'The dog swims through the river.'

The dataset was distributed over two annotation forms, and each annotation form was annotated by two native speakers. The annotators

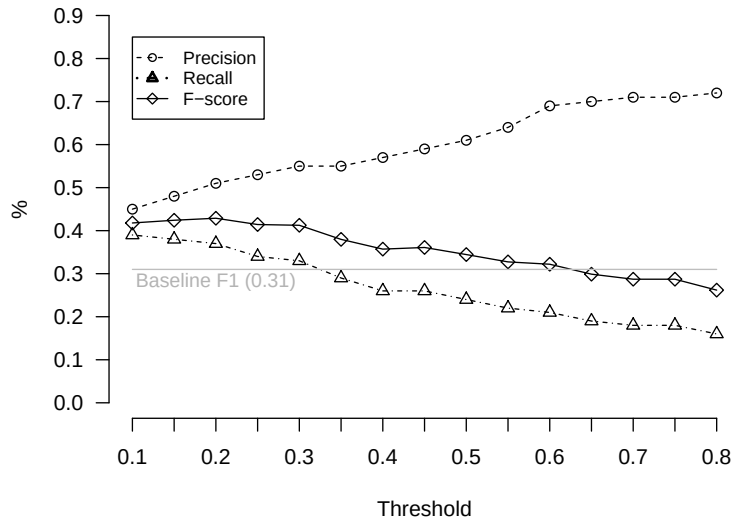


Figure 1: Trade-off between precision and recall across thresholds

had a background in linguistics or computational linguistics. They described the annotation as difficult to perform. This was also reflected by inter-annotator agreement; we observed fair agreement, Fleiss' $\kappa = 0.31$ (Landis and Koch 1977).

4.2.3 Results and discussion

Figure 1 presents the results when predicting slot correspondences, as measured by precision, recall and the harmonic F-score when comparing the system output to the human ratings. True positives were obtained if the system selected the same slot correspondence for a given slot that the human raters had selected. Since a variable threshold was applied, we find a trade-off between precision and recall. As expected, precision improves with higher thresholds, but this comes at the cost of lower recall. The F-score decreases with an increasing threshold, with a local maximum around a threshold of 0.2. With threshold values >0.6 the F-score drops below the baseline.

Overall, the system manages to predict correspondences between syntactic subcategorization slots to a fair degree of success. Our hypothesis that correspondence between subcategorization slots can be predicted by distributional semantic similarity has thus been confirmed. Then again, the success was not as high as we initially expected. We assume that this is due to the difficulty of the task, as indicated by the low inter-annotator agreement.

Since the annotators gave detailed comments after the annotation was completed, we detected theoretical problems which also apply to the automatic matching process. For example, the pair (17a)/(17b) for the verb *kleben* (*to stick/glue*) exemplifies a syntactic transfer of the theme argument *Zettel* (*note*), which is realized as the accusative object of the PV in (17a) and as the subject of the BV in (17b). The system failed to predict this transfer. This can be attributed to the fact that *kleben* can undergo a causative/inchoative alternation (Levin 1993), as exemplied by (17b)/(17c). We can observe a one-to-many match here. This is a problem which is hard to solve with our approach because the correspondence of PV–BV slots interferes with a slot correspondence among different uses of the BV.

(17) a. *Gerda KLEBT den Zettel an die Tür AN.*
Gerda STICKS the note on the door PRT_{an}
'Gerda sticks the note on the door.'

b. *Der Zettel KLEBT an der Tür.*
the note STICKS at the door
'The note sticks to the door.'

c. *Gerda KLEBT den Zettel an die Tür.*
Gerda STICKS the note at the door
'Gerda sticks the note on the door.'

Finally, we found that many of the feature vectors were extremely sparse, such as the vector of the PP headed by an_{dat} for the verb *an|heften* in Table 4. The sparsity problem could be remedied by reducing the number of dimensions, e.g, by applying some kind of abstraction over the head nouns. For example, the concepts of *Tür* (*door*) and *Kirchentür* (*church door*) are strongly related and could be merged into one dimension of the feature vector. The same holds for the concepts of *Pinnwand* (*pin board*), *Wand* (*wall*) and *Tafel* (*blackboard*). We suspect that with a certain level of abstraction over such concepts, the vectors would be more reliable. For this reason, we used generalization techniques in the following experiment.

4.3 *Experiment 3: Modeling distributional transfer*

In Section 2, we argued for a distributional assessment for predicting the degrees of compositionality for German PVs. We hypothesized that

the more compositional the PVs are, the more similar a PV and a BV are in their meanings and the more similar are their distributional properties. In the following, we suggest two types of distributional models in order to assess PV compositionality in a distributional manner:

1. *Window models*: If PVs occur in similar lexical contexts as their BVs, they are distributionally similar, which is taken as an indicator that the PVs are semantically similar to their BVs, hence highly compositional. In contrast, distributional distance should indicate lexical dissimilarity and thus low compositionality.
2. *Syntactic subcategorization models*: This approach models syntactic transfer: If PV subcategorization slots can be strongly mapped to subcategorization slots of their BVs, this indicates strong compositionality. The model thus integrates the prediction of slot correspondences between PVs and their BVs that was verified in the previous section.

The first option, *window models*, is conceptually very simple, since it compares unsorted local contexts. It does however not exploit the fact that local co-occurring words can be distinguished by their syntactic functions. Then again, window-based models accumulate an evidence mass which is proportionate to window size. One might suspect that this advantage in evidence mass comes at the cost of degraded quality, since windows represent bags of words.

The second option models the syntactic transfer and is thus theoretically more appealing because it distinguishes between context words according to their syntactic functions. Our hypothesis is that the degree of predicted associative strength of syntactic transfer represents an indicator of semantic transparency. If the complements of a PV strongly correspond to any complement of its BV, the PV is regarded as highly compositional, even if the PV complements are *not* realized as the same syntactic argument types, as long as a relation between these two subcategorization slots can be established. Conversely, if only a weak correspondence between the PV complements and the BV complements can be established, this is an indicator of low compositionality.

Our second approach is novel and exploits fine-grained syntactic transfer information, which is not accessible within a window-based approach. At the same time, it preserves an essential part of the in-

formation contained in context windows, since the head nouns within subcategorization frames typically appear in the local context.

The syntactic approach may however suffer from a practical problem, i.e., data sparseness. While in the case of window information every instance of a verb has *2*n* words in the local context, in the transfer approach each verb instance has just as many co-occurring words as it has subcategorization slots. To compensate for this inevitable data sparseness, we employed the lexical taxonomy *GermaNet* (Hamp and Feldweg 1997) and *Singular Value Decomposition (SVD)* to generalize over individual complement heads. Dimensionality reduction techniques have proven effective in previous distributional semantics tasks (e.g., Joanis *et al.* 2008, Brody and Elhada 2010, Ó Séaghdha 2010, Guo and Diab 2011, Bullinaria and Levy 2012, Turney 2012).

1. *GermaNet (GN)* (Hamp and Feldweg 1997) is the German version of WordNet (Fellbaum 1998). We used the n^{th} topmost taxonomy levels in the GermaNet hierarchy as generalizations of head nouns. In the case of multiple inheritance, the counts of a subordinate node were distributed over the superordinated nodes.
2. *Singular Value Decomposition (SVD)*: We used the DISSECT tool (Dinu *et al.* 2013) to apply singular value decomposition to the vectors of complement head nouns in order to reduce the dimensionality of the vector space.

GermaNet is a knowledge-driven way of mapping concepts to more general concepts; SVD learns abstract latent dimensions automatically.

4.3.1 Experimental setup

Window Model: For the assessment of PV compositionality based on windows we used a word vector space model (Sahlgren 2006; Turney and Pantel 2010). The experiment replicates and extends an approach presented in Bott and Schulte im Walde (2014b), where we demonstrated the reliability of window-based models to predict PV compositionality and assessed the effect of target frequency, ambiguity, and lemma restoration. For each target PV, we constructed a vector space with s_l dimensions, where s_l was the size of the vocabulary as extracted from a lemmatized corpus. The vector components represened co-occurrence counts in local context, which was defined as a window of n words to the left and to the right of the target PV.

In our experiment with window-based models, words were lemmatized, but no dimensionality reduction was applied. Since PVs may occur in syntactically separated paradigms (i.e., the particle separated from the verb), but lemmatizers are blind to syntactic dependencies, we applied lemma correction: If we found a verb particle which the parser resolved as directly depending on a verb, we concatenated the particle with the verb lemma in order to derive the lemma of the PV. Our models vary (a) in the size of the context window, (b) by (not) applying term-weighting, and (c) by using all context words or only content words as vector dimensions. Windows did not go beyond sentence boundaries, because our corpora were sentence-shuffled for copyright reasons. The semantic similarity, which is taken as the associative strength of a PV–BV pair $<pv, bv>$ was calculated as the cosine between the vectors for pv and bv.

Syntactic Subcategorization Model: The rationale behind the use of syntactic slot correspondence to predict the degree of PV–BV compositionality is that we only try to match those semantic arguments which correspond to each other. This requires two steps: first, detecting the best matching slots in PV–BV pairs; second, determining their average distributional similarity. Relying on the five most frequent subcategorization frames, we first selected the best matching BV slot for each PV complement slot, as described in 4.2, and then calculated the associative strength as_{pv}^{bv} between a PV–BV pair $<pv, bv>$ as the average cosine score over the best matches for all PV slots and the best matches for all BV slots. The associative strength as_{pv}^{bv} is taken as a measure of the correspondence of PV–BV complement slots and their realization of the same semantic arguments. We thus take the strength to predict the degree of PV compositionality. To account for possible null correspondences in argument incorporation and argument extension cases, we applied a variable threshold on the cosine distance ($t = 0.1/0.2/0.3$). If the best matching BV complement slot of a PV complement slot had a cosine score below this threshold, it was not taken into account. $t = 0$ refers to setting no threshold.

4.3.2 Vector weighting and Generalization

Not all context words are equally predictive for lexical distributional models: Some words tend to occur frequently across many contexts, which makes them bad predictors. We thus leveraged information

which stems from words that occur in specific contexts and were expected to represent salient predictors. To this end, we used *local mutual information* (LMI, Evert 2004) as a vector weighting method and test if term weighting has an effect on the prediction quality. To filter out the distortion introduced by non-content words, we used window models which only contain context information corresponding to nouns, verbs and adjectives. To address the second representation issue, data sparseness in syntactic subcategorization models, we applied GermaNet and SVD as generalizations.

4.3.3 Corpora

In order to estimate the effect that the amount of data has on the prediction quality, we compare vector spaces from two differently sized corpora. As in the previous two experiments, we used the dependency-parsed SdeWaC corpus with ≈880 million words. In comparison, we used the DECOW14[9] corpus (Schäfer and Bildhauer 2012) with ≈20 billion words. The DECOW14 data was pre-processed and dependency-parsed with a toolchain presented in Björkelund *et al.* (2013): Their pipeline used the graph-based MATE dependency parser (Bohnet 2010), which was also used for the preprocessing of the SdeWaC corpus. For morphological analysis MarMoT (Müller *et al.* 2013) and SMOR (Schmid *et al.* 2004) were applied.

4.3.4 Gold standards

We evaluated our models against three gold standards (GSs). Each of them contains PVs across different particles and was annotated by humans for the degree of compositionality:

1. ***GS1***: A gold standard collected by Hartmann (2008), consisting of 99 randomly selected PVs across 11 particles, balanced over 8 frequency ranges and judged by 4 experts on a scale from 0 to 10.
2. ***GS2***: A gold standard of 354 randomly selected PVs across the same 11 particles, balanced over 3 frequency ranges while taking the frequencies from 3 corpora into account. Ratings were collected with Amazon Mechanical Turk on a scale from 1 to 7.
3. ***GS3***: A cleaned subset of 150 PVs from GS2, after removing the most frequent and infrequent PVs as well as prefix verbs.[10]

[9] `http://corporafromtheweb.org/decow14/`

[10] Some verbs such as *um|fahren* do exists as both PVs and prefix verbs.

We compared the rankings of the system-derived PV–BV cosine scores against the human ratings, using Spearman's rank-order correlation coefficient ρ (Siegel and Castellan 1988).

4.3.5 Results and discussion

Window Model: Figure 2 presents the general results for different window sizes and across the three gold standards. All of the ρ scores correspond to very high levels of statistical significance ($p<0.005$). The results tend to improve slightly with increasing window sizes. For very large windows, especially for sizes 15 and above, the results remain at the same level, except for GS1 which slightly drops. This is not surprising since windows were cut at sentence boundaries which in practice makes the sentence length the upper bound for the window size.

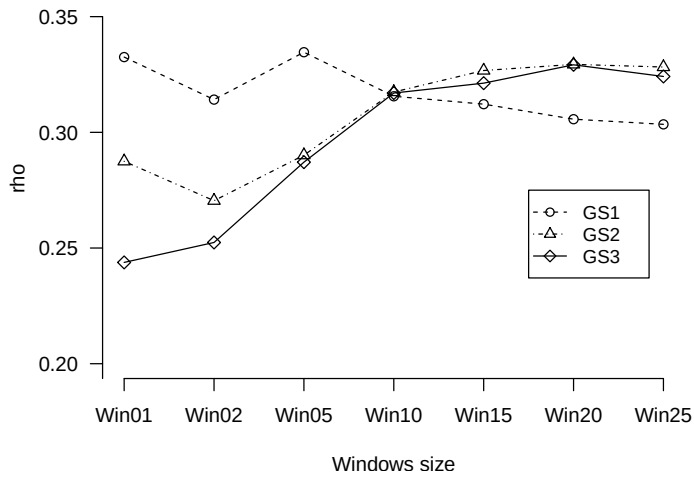


Figure 2: Results for differently sized window models across the three gold standards. The models rely on content words and use LMI weighting

Results for GS1 based on the SdeWaC vs. the DECOW14 corpus are shown in Figure 3. The performance of the two groups of models is largely comparable, and no clear advantage of one over the other is observable. Given that DECOW is considerably larger than SdeWaC, we take this as evidence that window models are relatively robust against data sparseness.

Figure 4 compares models that use raw frequency counts for all context words with using only content words, combined with LMI weighting. Clearly, the latter type of model leads to far better results.

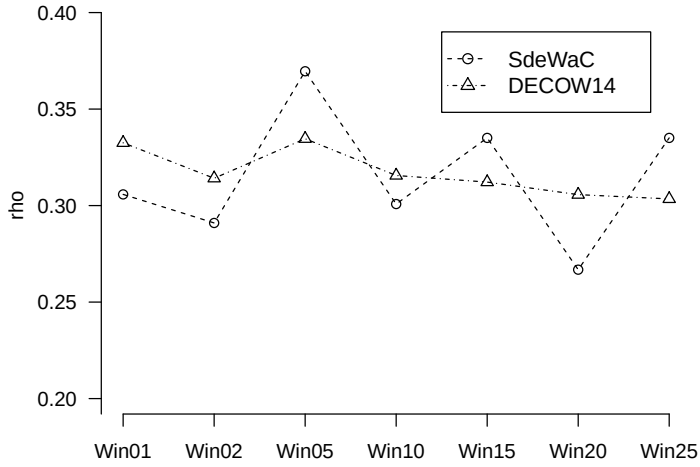


Figure 3: Results for GS1 with window models extracted from two different corpora: SdeWaC and DECOW14. The models rely on content words and use LMI weighting

Syntactic Subcategorization Model: As for models that take syntactic transfer strength into account, Figure 5 shows the overall results for subcategorization models with a threshold of $t = 0.3$. The first set of bars represents the best window model as a point of comparison, i.e., using a window of 20 words, reduced to content words, and with LMI weighting. The following groups of bars represent syntactic transfer models with raw frequency counts, LMI weighting, GermaNet generalizations (gn.lv$x$) and SVD (svd_dim) dimensionality reductions.

Two observations can be made: firstly, none of the syntactic models reaches the level of performance of the window-based models. Second, the high-dimensional models based on raw frequency counts and LMI perform much worse than the models which apply generalization techniques. So, contrary to the window-based models, applying LMI weighting does not improve the predictions. But generalizations boost the quality of the predictions in many conditions.

The fact that the concentration of evidence mass through generalization by GermaNet and SVD greatly benefits the results suggests that the major problem of the syntactic subcategorization approach is data sparseness. The use of GermaNet generalizations already tends to improve the performance, although not consistently. But the use of such taxonomy-based generalizations is clearly limited by the fact that taxonomies notoriously lack coverage and, in the frequent case of semantic ambiguity, are not able to provide reliable estimates on

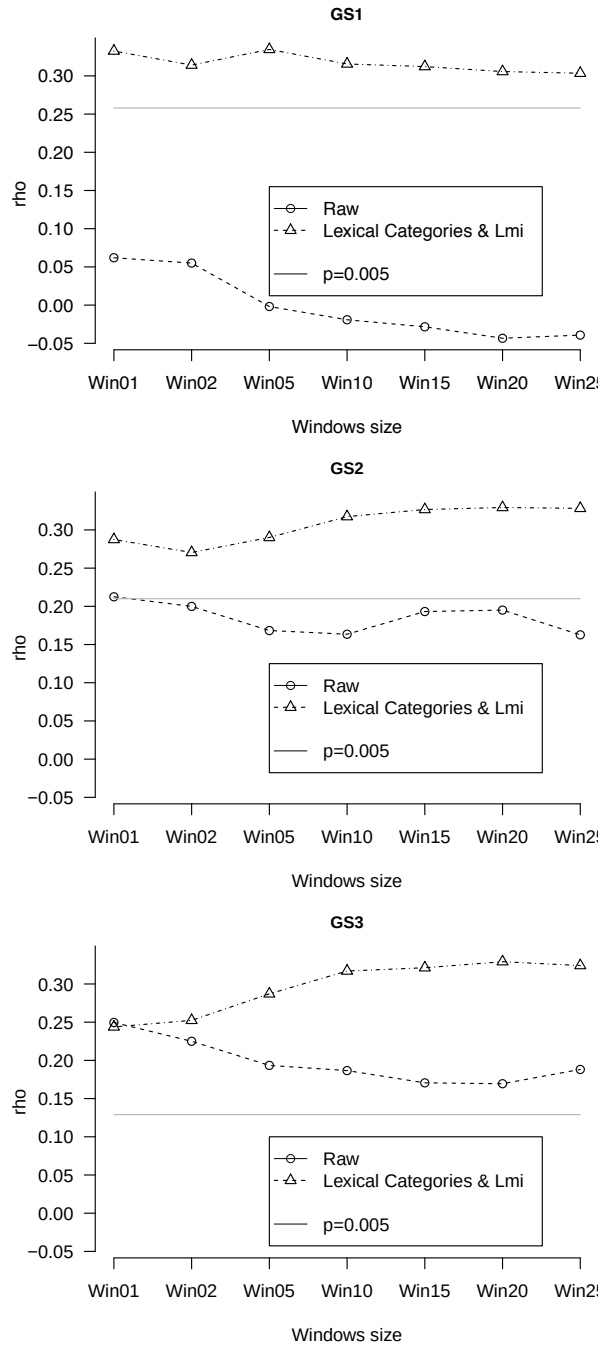


Figure 4:
Results for raw frequency models vs. models with content words and LMI weighting

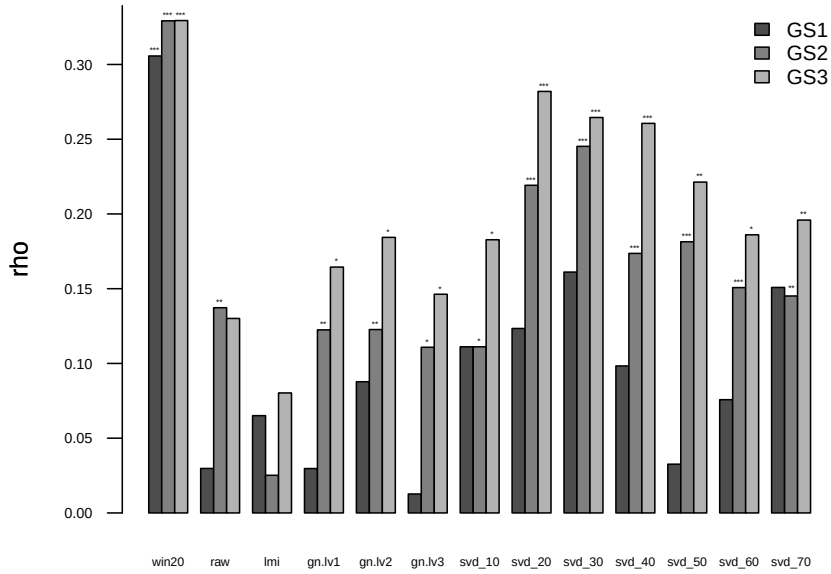


Figure 5: Results across gold standards, for t = 0.3 (*** p<0.001, ** p<0.01, * p<0.05)

the probabilities of different word readings. A major boost in performance can be observed with the use of SVD, which does not run into coverage problems. The best SVD results are obtained in the range of twenty dimensions (svd_20), which seems to be the best equilibrium between the concentration of evidence mass and over-generalization.

A similar effect can also be observed for GermaNet generalizations: the highest level of distinction in the taxonomy (gn.lv1) is too general to be useful while the third (gn.lv3) is too specific; the second level of the taxonomy (gn.lv2) appears to be the best compromise.

The assumption that data sparseness plays a major role in the performance of the syntactic subcategorization models is also backed up by a comparison between models extracted from our differently sized corpora, as presented in Figure 6. It is important to keep in mind that the SdeWaC corpus itself is not a small corpus, but the use of the much larger DECOW14 leads to better results in most cases. This stands in sharp contrast to the window-based models which, as we have seen above, apparently do not improve with the larger corpus and do not run into data sparseness problems.

As discussed earlier, we suspected that information stemming from window models provides semantic evidence of a somewhat degraded quality. For this reason, the evidence extracted from syntactic

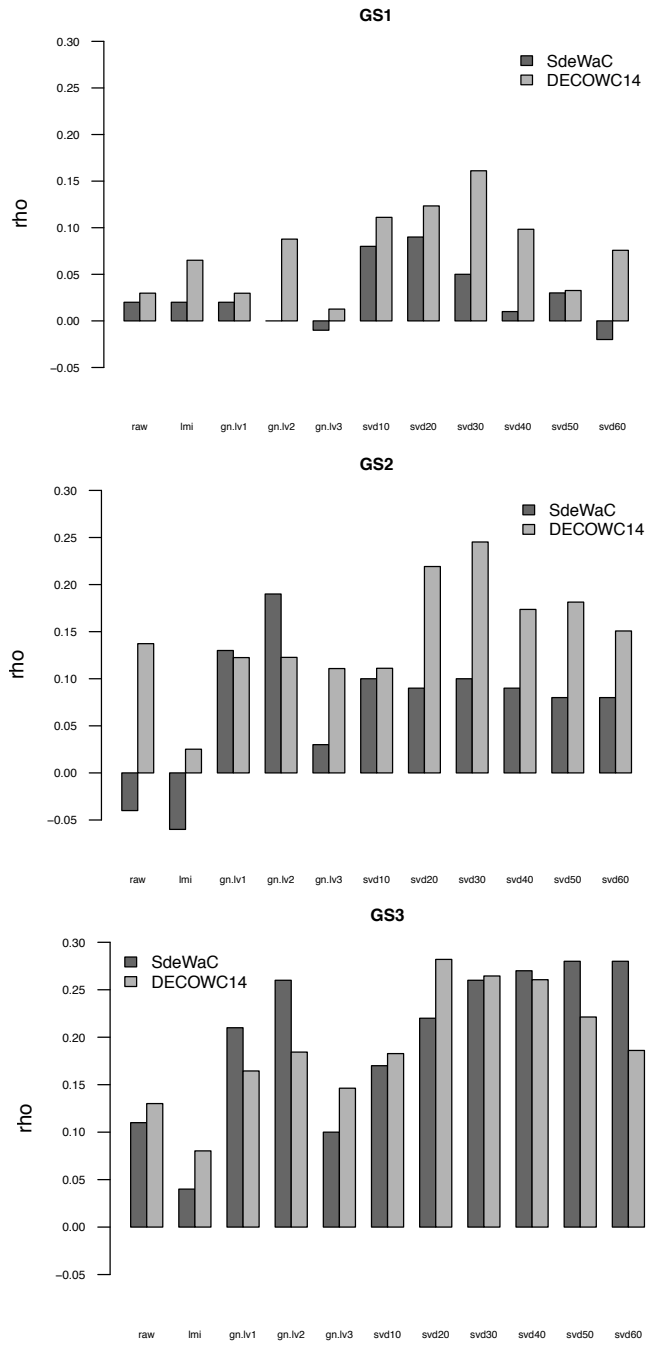


Figure 6: Results for syntactic models extracted from two different corpora: SdeWaC vs. DECOW14

slot fillers should in theory be qualitatively better. But if we assume that information stemming from the argument grid and the heads of syntactic relations is qualitatively more valuable information for our task, we should expect that larger window sizes do not predict compositionality as well as small or medium-sized windows, since small windows tend to contain more concentrated material from arguments than very large windows. What we found in Figure 2, however, is that in general large windows lead to a better performance than small windows. This strongly suggests that words from the general context, which are not necessarily syntactically linked to our target verbs in a direct way, are also very valuable predictors for the semantic similarity between PV–BV pairs and, thus, their level of compositionality. This also means that building our theoretical considerations about the matching of argument slots between PV–BV pairs does not outweight the larger mass of unsorted evidence contained in the window models.

A further problem of our syntax-aware approach is revealed if we look at Figure 7, which compares the prediction results across thresholds t. We can see that a threshold of 0.2 or 0.3 often leads to a slightly better performance than 0.1 or no threshold, but no globally optimal value for t can be established. If the threshold is set too low, many non-correspondences are interpreted as semantic links (false positives). If the threshold is set too high, many semantic links are discarded (false negatives). There seems to be no optimal point of equilibrium between the filtering of false positives and false negatives. A dynamic threshold for individual PV–BV pairs and the average cosine distances of a target slot to all given complementary candidate slots would be beneficial, but at present we see no way to compute this reliably.

Finally, and with respect to the last problem, our syntax-based approach somewhat naively neglects the possibility of one-to-many and many-to-one correspondences between subcategorization slots, and always tries to establish a one-to-one link. In reality, however, many subcategorization slots with more than one correct correspondence can be found. For example, the PV–BV pair *leuchten/an|leuchten* as in example (10) happens to be a classification outlier in many of the syntax-based prediction models. The subject slot (SB) of the BV *leuchten* (e.g., *Lampe (lamp)*) is usually matched to a PP subcategorization slot of the PV *an|leuchten* headed by the preposition *mit*, which requires the dative case (e.g., *mit der Lampe (with the lamp)*). Our sys-

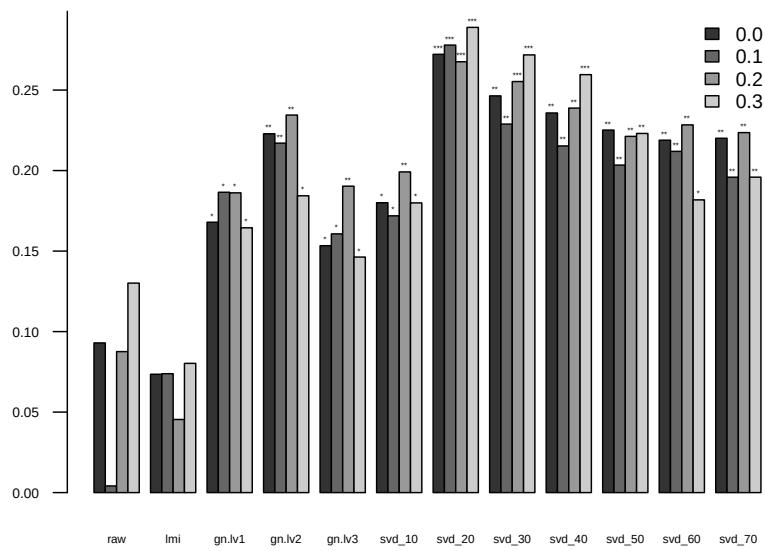


Figure 7: Results across thresholds, GS3 (*** p<0.001, ** p<0.01, * p<0.05)

tem computed the following two slots for *leuchten* which receive high cosine values in correspondence to the PP `mit-dat` slot of *an|leuchten*.

```
anleuchten-mit-dat vs leuchten-SB:     0.8931
anleuchten-mit-dat vs leuchten-in-dat: 0.6386
```

One slot is the subject (SB), as expected, and the second is a PP headed by the preposition *in* and the dative case. The latter option represents a linguistically plausible complement of *leuchten* indicating the location where the illumination takes place (e.g., *leuchtet in dem Raum (shines in the room)*), but without semantic correspondence to the target PV slot. A possible remedy for our prediction model could be to include an estimation about how many links have to be established, but this is not a trivial problem in itself and will not be pursued here.

In sum, we provided empirical evidence for hypothesis H3: we found that both window models and syntactic models that are sensitive to subcategorization frame transfers can be used to predict degrees of PV compositionality. Window-based models perform better, even though they are conceptually and computationally simpler. The worse performance of the syntactic models is presumably due to data sparseness and underlying linguistic problems which are difficult to solve computationally.

5 CONCLUSION

At the beginning of this article, we hypothesized that for PVs that are not fully lexicalized there are groups of BVs which undergo the same semantic derivation when they combine with the same particle type, and that the semantic transfer patterns are paralleled by syntactic transfer patterns. We further hypothesized that syntactic transfer between pairs of PVs and BVs, as well as the degree of PV compositionality, can be predicted with distributional methods.

Our first experiment in Section 4.1 addressed the hypothesis that particle meaning and semantic derivation are closely related. We found evidence that there are groups of PVs which share the same semantic transfer patterns and also the same syntactic transfer patterns. This shows that the PVs in the same semantic classes (i) are semantically coherent, (ii) share semantically coherent BVs and the same particle senses, and (iii) undergo parallel shifts regarding syntactic and semantic properties. We thus contributed both to the theoretical understanding and to an empirical verification of German PV composition at the syntax-semantics interface.

Our second experiment in Section 4.2 addressed the empirical prediction of PV–BV syntactic subcategorization transfer, which we argued is necessary to integrate into a prediction of PV compositionality from a theoretical point of view. While modeling slot correspondences in the syntactic transfer was challenging for humans and suffered from severe data sparseness, we verified our distributional approach using hard and soft cluster analyses.

Finally, our third experiment in Section 4.3 integrated the idea of slot correspondence into a syntactic transfer model of PV compositionality, and compared the syntactic model against window models. Although the syntactic transfer approach is much more elaborate and theoretically well-founded, it could not outperform the conceptually simpler window-based approach. We argued that local windows contain information which is useful in the prediction of semantic similarity between PV–BV pairs, and which apparently captures aspects of the verb meanings that the syntactic complements are missing. The window-based approach also proved more robust to data sparseness. Overall, we found that both models can be used to predict degrees of PV compositionality, and the comparison between the two approaches allowed important theoretical insights: many of the misclassifications

produced by the syntax-based models could be traced to underlying linguistic problems, the complexity of which makes computational analysis infeasable given the available resources.

ACKNOWLEDGMENTS

The research was supported by the DFG Research Grant SCHU 2580/2 (Stefan Bott) and the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde). We thank the anonymous reviewers and our colleagues Jeremy Barnes and Diego Frassinelli for their helpful feedback, and our annotators for the tedious manual annotations of particle verb syntax and semantics.

REFERENCES

Nadine ALDINGER (2004), Towards a Dynamic Lexicon: Predicting the Syntactic Argument Structure of Complex Verbs, in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Timothy BALDWIN, Colin BANNARD, Takaaki TANAKA, and Dominic WIDDOWS (2003), An Empirical Model of Multiword Expression Decomposability, in *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 89–96, Sapporo, Japan.

Collin BANNARD (2005), Learning about the Meaning of Verb–Particle Constructions from Corpora, *Computer Speech and Language*, 19:467–478.

Anders BJÖRKELUND, Özlem ÇETINOĞLU, Richárd FARKAS, Thomas MÜLLER, and Wolfgang SEEKER (2013), (Re)ranking Meets Morphosyntax: State-of-the-art Results from the SPMRL 2013 Shared Task, in *Proceedings of the 4th Workshop on Statistical Parsing of Morphologically-Rich Languages*, pp. 135–145, Seattle, WA, USA.

David BLEI, Andrew NG, and Michael JORDAN (2003), Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3:993–1022.

Bernd BOHNET (2010), Top Accuracy and Fast Dependency Parsing is not a Contradiction, in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 89–97, Beijing, China.

Stefan BOTT and Sabine SCHULTE IM WALDE (2014a), Modelling Regular Subcategorization Changes in German Particle Verbs, in *Proceedings of the 1st Workshop on Computational Approaches to Compound Analysis*, pp. 1–10, Dublin, Ireland.

Stefan BOTT and Sabine SCHULTE IM WALDE (2014b), Optimizing a Distributional Semantic Model for the Prediction of German Particle Verb

Compositionality, in *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pp. 509–516, Reykjavik, Iceland.

Stefan BOTT and Sabine SCHULTE IM WALDE (2014c), Syntactic Transfer Patterns of German Particle Verbs and their Impact on Lexical Semantics, in *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics*, pp. 182–192, Dublin, Ireland.

Stefan BOTT and Sabine SCHULTE IM WALDE (2015), Exploiting Fine-grained Syntactic Transfer Features to Predict the Compositionality of German Particle Verbs, in *Proceedings of the 11th Conference on Computational Semantics*, pp. 34–39, London, UK.

Samuel BRODY and Noemie ELHADA (2010), An Unsupervised Aspect-Sentiment Model for Online Reviews, in *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 804–812, Los Angeles, CA, USA.

John A. BULLINARIA and Joseph P. LEVY (2012), Extracting Semantic Representations from Word Co-Occurrence Statistics: Stop-Lists, Stemming, and SVD, *Behavior Research Methods*, 44:890–907.

Fabienne CAP, Manju NIRMAL, Marion WELLER, and Sabine SCHULTE IM WALDE (2015), How to Account for Idiomatic German Support Verb Constructions in Statistical Machine Translation, in *Proceedings of the 11th Workshop on Multiword Expressions*, pp. 19–28, Denver, Colorado, USA.

Kostadin CHOLAKOV and Valia KORDONI (2014), Better Statistical Machine Translation through Linguistic Treatment of Phrasal Verbs, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 196–201, Doha, Qatar.

Paul COOK and Suzanne STEVENSON (2006), Classifying Particle Semantics in English Verb-Particle Constructions, in *Proceedings of the ACL/COLING Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pp. 45–53, Sydney, Australia.

Nicole DEHÉ, Ray JACKENDOFF, Andrew MCINTYRE, and Silke URBAN (2002), Introduction, in Nicole DEHÉ, Ray JACKENDOFF, Andrew MCINTYRE, and Silke URBAN, editors, *Verb-Particle Explorations*, Interface Explorations, pp. 1–20, Mouton de Gruyter, Berlin, New York.

Georgiana DINU, Nghia THE PHAM, and Marco BARONI (2013), DISSECT – DIStributional SEmantics Composition Toolkit, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 31–36, Sofia, Bulgaria.

Stefan EVERT (2004), The Statistical Analysis of Morphosyntactic Distributions, in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1539–1542, Lisbon, Portugal.

Gertrud FAAß and Kerstin ECKART (2013), SdeWaC – A Corpus of Parsable Sentences from the Web, in *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, pp. 61–68, Darmstadt, Germany.

Christiane FELLBAUM, editor (1998), *WordNet – An Electronic Lexical Database*, Language, Speech, and Communication, MIT Press, Cambridge, MA.

John R. FIRTH (1957), *Papers in Linguistics 1934–51*, Longmans, London, UK.

Wolfgang FLEISCHER and Irmhild BARZ (2012), *Wortbildung der deutschen Gegenwartssprache*, Walter de Gruyter, 4th edition.

Joseph L. FLEISS (1971), Measuring Nominal Scale Agreement among Many Raters, *Psychological Bulletin*, 76(5):378–382.

Weiwei GUO and Mona DIAB (2011), Semantic Topic Models: Combining Word Distributional Statistics and Dictionary Definitions, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 552–561, Edinburgh, UK.

Birgit HAMP and Helmut FELDWEG (1997), GermaNet – A Lexical-Semantic Net for German, in *Proceedings of the ACL Workshop on Automatic Information Extraction and Building Lexical Semantic Resources for NLP Applications*, pp. 9–15, Madrid, Spain.

Zellig HARRIS (1954), Distributional Structure, *Word*, 10(23):146–162.

Silvana HARTMANN (2008), Einfluss syntaktischer und semantischer Subkategorisierung auf die Kompositionalität von Partikelverben, Studienarbeit, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Boris HASELBACH (2011), Deconstructing the Meaning of the German Temporal Verb Particle *nach* at the Syntax-Semantics Interface, in *Proceedings of Generative Grammar in Geneva*, pp. 71–92, Geneva, Switzerland.

Lawrence HUBERT and Phipps ARABIE (1985), Comparing Partitions, *Journal of Classification*, 2:193–218.

Eric JOANIS, Suzanne STEVENSON, and David JAMES (2008), A General Feature Space for Automatic Verb Classification, *Natural Language Engineering*, 14(3):337–367.

Fritz KLICHE (2011), Semantic Variants of German Particle Verbs with *ab-*, *Leuvense Bijdragen*, 97:3–27.

Maximilian KÖPER and Sabine SCHULTE IM WALDE (2017), Complex Verbs are Different: Exploring the Visual Modality in Multi-Modal Models to Predict Compositionality, in *Proceedings of the 13th Workshop on Multiword Expressions*, pp. 200–206, Valencia, Spain.

Maximilian KÖPER and Sabine SCHULTE IM WALDE (2018), Analogies in Complex Verb Meaning Shifts: The Effect of Affect in Semantic Similarity

Models, in *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, LA, USA, to appear.

Anna KORHONEN, Yuval KRYMOLOWSKI, and Zvika MARX (2003), Clustering Polysemic Subcategorization Frame Distributions Semantically, in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 64–71, Sapporo, Japan.

Natalie KÜHNER and Sabine SCHULTE IM WALDE (2010), Determining the Degree of Compositionality of German Particle Verbs by Clustering Approaches, in *Proceedings of the 10th Conference on Natural Language Processing*, pp. 47–56, Saarbrücken, Germany.

Thomas K. LANDAUER and Susan T. DUMAIS (1997), A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge, *Psychological Review*, 104(2):211–240.

J. Richard LANDIS and Gary G. KOCH (1977), The Measurement of Observer Agreement for Categorical Data, *Biometrics*, 33:159–174.

Andrea LECHLER and Antje ROßDEUTSCHER (2009), German Particle Verbs with *auf*. Reconstructing their Composition in a DRT-based Framework, *Linguistische Berichte*, 220:439–478.

Beth LEVIN (1993), *English Verb Classes and Alternations*, The University of Chicago Press.

Anke LÜDELING (2001), *On German Particle Verbs and Similar Constructions in German*, Dissertations in Linguistics, CSLI Publications, Stanford, CA.

Christopher D. MANNING, Prabhakar RAGHAVAN, and Hinrich SCHÜTZE (2008), *Introduction to Information Retrieval*, Cambridge University Press.

Diana MCCARTHY, Bill KELLER, and John CARROLL (2003), Detecting a Continuum of Compositionality in Phrasal Verbs, in *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 73–80, Sapporo, Japan.

Andrew MCINTYRE (2007), Particle Verbs and Argument Structure, *Language and Linguistics Compass*, 1(4):350–397.

Paola MERLO and Suzanne STEVENSON (2001), Automatic Verb Classification Based on Statistical Distributions of Argument Structure, *Computational Linguistics*, 27(3):373–408.

Stefan MÜLLER (2002), Syntax or Morphology: German Particle Verbs Revisited, in Nicole DEHÉ, Ray JACKENDOFF, Andrew MCINTYRE, and Silke URBAN, editors, *Verb-Particle Explorations*, Interface Explorations, pp. 119–140, Mouton de Gruyter, Berlin, New York.

Stefan MÜLLER (2003), Solving the Bracketing Paradox: An Analysis of the Morphology of German Particle Verbs, *Journal of Linguistics*, 39(2):275–325.

Thomas MÜLLER, Helmut SCHMID, and Hinrich SCHÜTZE (2013), Efficient Higher-Order CRFs for Morphological Tagging, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 322–332, Seattle, WA, USA.

Diarmuid Ó SÉAGHDHA (2010), Latent Variable Models of Selectional Preference, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 435–444, Uppsala, Sweden.

Susan OLSEN (1997), Prädikative Argumente syntaktischer und lexikalischer Köpfe—Zum Status der Partikelverben im Deutschen und Englischen, *Folia Linguistica*, 31(3–4):301–330.

Sebastian PADÓ and Mirella LAPATA (2007), Dependency-based Construction of Semantic Space Models, *Computational Linguistics*, 33(2):161–199.

William M. RAND (1971), Objective Criteria for the Evaluation of Clustering Methods, *Journal of the American Statistical Association*, 66(336):846–850.

Mats ROOTH (1998), Two-Dimensional Clusters in Grammatical Relations, in *Inducing Lexicons with the EM Algorithm*, AIMS Report 4(3), Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Mats ROOTH, Stefan RIEZLER, Detlef PRESCHER, Glenn CARROLL, and Franz BEIL (1999), Inducing a Semantically Annotated Lexicon via EM-Based Clustering, in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 104–111, Maryland, MD.

Stefan RÜD (2012), *Untersuchung der distributionellen Eigenschaften der Lesarten der Partikel 'auf' mittels Clustering-Methoden*, Master's thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Magnus SAHLGREN (2005), An Introduction to Random Indexing, in *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*, volume 5.

Magnus SAHLGREN (2006), *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*, Ph.D. thesis, Stockholm University.

Roland SCHÄFER and Felix BILDHAUER (2012), Building Large Corpora from the Web Using a New Efficient Tool Chain, in *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pp. 486–493, ELRA, Istanbul, Turkey.

Silke SCHEIBLE, Sabine SCHULTE IM WALDE, Marion WELLER, and Max KISSELEW (2013), A Compact but Linguistically Detailed Database for German Verb Subcategorisation Relying on Dependency Parses from a Web Corpus: Tool, Guidelines and Resource, in *Proceedings of the 8th Web as Corpus Workshop*, pp. 63–72, Lancaster, UK.

Helmut SCHMID, Arne FITSCHEN, and Ulrich HEID (2004), SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection, in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1263–1266.

Sabine SCHULTE IM WALDE (2000), Clustering Verbs Semantically According to their Alternation Behaviour, in *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 747–753, Saarbrücken, Germany.

Sabine SCHULTE IM WALDE (2004), Identification, Quantitative Description, and Preliminary Distributional Analysis of German Particle Verbs, in *Proceedings of the COLING Workshop on Enhancing and Using Electronic Dictionaries*, pp. 85–88, Geneva, Switzerland.

Sabine SCHULTE IM WALDE (2005), Exploring Features to Identify Semantic Nearest Neighbours: A Case Study on German Particle Verbs, in *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pp. 608–614, Borovets, Bulgaria.

Sabine SCHULTE IM WALDE (2006), Experiments on the Automatic Induction of German Semantic Verb Classes, *Computational Linguistics*, 32(2):159–194.

Sidney SIEGEL and N. John CASTELLAN (1988), *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, Boston, MA.

Sylvia SPRINGORUM (2011), DRT-based Analysis of the German Verb Particle *an*, *Leuvense Bijdragen*, 97:80–105.

Sylvia SPRINGORUM, Sabine SCHULTE IM WALDE, and Antje ROßDEUTSCHER (2012), Automatic Classification of German *an* Particle Verbs, in *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pp. 73–80, Istanbul, Turkey.

Sylvia SPRINGORUM, Sabine SCHULTE IM WALDE, and Antje ROßDEUTSCHER (2013a), Sentence Generation and Compositionality of Systematic Neologisms of German Particle Verbs, Talk at the 5th Conference on Quantitative Investigations in Theoretical Linguistics.

Sylvia SPRINGORUM, Jason UTT, and Sabine SCHULTE IM WALDE (2013b), Regular Meaning Shifts in German Particle Verbs: A Case Study, in *Proceedings of the 10th International Conference on Computational Semantics*, pp. 228–239, Potsdam, Germany.

Barbara STIEBELS (1996), *Lexikalische Argumente und Adjunkte. Zum semantischen Beitrag von verbalen Präfixen und Partikeln*, Akademie Verlag, Berlin.

Barbara STIEBELS and Dieter WUNDERLICH (1994), Morphology Feeds Syntax: The Case of Particle Verbs, *Linguistics*, 32:913–968.

Peter D. TURNEY (2012), Domain and Functions: A Dual-Space Model of Semantic Relations and Compositions, *Journal of Artificial Intelligence Research*, 44:533–585.

Peter D. TURNEY and Patrick PANTEL (2010), From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research*, 37:141–188.

Marion WELLER, Fabienne CAP, Stefan MÜLLER, Sabine SCHULTE IM WALDE, and Alexander FRASER (2014), Distinguishing Degrees of Compositionality in Compound Splitting for Statistical Machine Translation, in *Proceedings of the 1st Workshop on Computational Approaches to Compound Analysis*, pp. 81–90, Dublin, Ireland.

Ian H. WITTEN and Eibe FRANK (2005), *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann.

Stefanie WULFF (2010), *Rethinking Idiomaticity: A Usage-Based Approach*, A&C Black.

Integrating LFG's binding theory with PCDRT

Mary Dalrymple[1]*, Dag T. T. Haug*[2]*, and John J. Lowe*[1]
[1] Faculty of Linguistics, Philology and Phonetics, University of Oxford
[2] Department of Linguistics and Scandinavian Studies, University of Oslo

ABSTRACT

We provide a formal model for the interaction of syntax and pragmatics in the interpretation of anaphoric binding constraints on personal and reflexive pronouns. We assume a dynamic semantics, where type *e* expressions introduce discourse referents, and contexts are assignments of individuals to discourse referents. We adopt the Partial Compositional Discourse Representation Theory (PCDRT) of Haug (2014b), whereby anaphoric resolution is modelled in terms of a pragmatically-established relation between discourse referents. We integrate PCDRT into the constraint-based grammatical framework of Lexical Functional Grammar (LFG), and show how it is possible to state syntactic constraints on the pragmatic resolution of singular and plural anaphora within this framework.

Keywords: anaphora, binding theory, Lexical Functional Grammar, Partial Compositional Discourse Representation Theory

1 INTRODUCTION

Pronouns are among the most frequently occurring words in many languages, including English, and speakers find no difficulty in using them and, for the most part, determining their reference in a particular context. However, formally analysing the constraints on the interpretation of pronouns in context is a complex matter. In part, this is due to the fact that the interpretation of pronominal reference involves two components of language which are usually considered separate: syntax and pragmatics. While syntax and semantics are widely

treated as part of a formal computational system which pairs linguistic form with linguistic meaning, pragmatic interpretation is widely treated as a distinct system which interprets linguistic meaning but which is separate from grammar itself. The interpretation of pronouns is one instance of a linguistic phenomenon which brings into question the sharp separation of pragmatics from the rest of grammar.[1] The Partial Compositional Discourse Representation Theory (PCDRT) of Haug (2014b) addresses this by modelling anaphoric resolution in terms of a pragmatically-established relation between discourse referents. This creates a clean separation of monotonic from non-monotonic aspects of interpretation, licensing an integration of pragmatics into grammatical modelling without undermining the monotonic nature of the rest of grammar.

In this paper we provide a formal model for the interaction of syntax and pragmatics in the interpretation of anaphoric binding constraints on personal and reflexive pronouns. We implement the model using PCDRT as the semantic framework, but the only information that needs to be available at the syntax-pragmatics interface is a function that takes anaphoric expressions to their antecedents. Our approach is therefore compatible with any theory that models anaphor-antecedent relationships in this way (as opposed to, say, equating anaphor and antecedent variables in the syntax).

Nevertheless, we believe there is considerable value in demonstrating how PCDRT can be integrated into the constraint-based grammatical framework of Lexical Functional Grammar (LFG). The most complex data that we consider in this paper – negative binding constraints on plural pronouns – have to our knowledge not been treated since the work of Berman and Hestvik (1997), which is cast in a framework where Government and Binding-style trees are rewritten as DRSs. This approach to the syntax-semantics interface is hardly in use today, whereas we are assuming a standard, lambda-driven approach to semantic composition. From a formal point of view, LFG and PCDRT naturally complement each other. LFG distinguishes different types of grammatical information and treats them as distinct levels of representation. Their combination produces a model in which syn-

[1] For a detailed discussion of other phenomena that point in this direction, see e.g. Chierchia (2004).

tactic, semantic and pragmatic constraints on pronoun resolution can be integrated while remaining distinct. This allows us to extend the empirical coverage of LFG, in particular by providing a complete and formally explicit account of negative binding constraints, which has not been available in previous work.

2 SYNTACTIC BINDING CONDITIONS

It is clear enough that at least to some extent, the resolution of pronominal anaphora is pragmatically based. Given the following sentence, the hearer cannot determine syntactically or semantically whether *he* refers to *Bertie* or to a different individual (available from the discourse or wider context); this determination can only be made pragmatically, based on the context of the utterance.

(1) Bertie knew that he wanted to leave.

Most importantly, the determination of a particular antecedent for a pronoun is not fixed: once it is made, it can be revised if the subsequent discourse provides additional information which contradicts the assignment made. This ability to update the relations between pronouns and antecedents shows that the relation is fundamentally pragmatic. This will be discussed in Section 5.

At the same time, pronominal binding is generally subject to syntactic constraints of various kinds, defined in terms of a superiority relation between the pronoun and its allowed and disallowed binders, and the syntactic domain in which a pronoun must or must not be bound. The classic binding theory of Chomsky (1981) defines the following constraints:

(2) Binding conditions according to Chomsky (1981, 188):

> **Principle A**: An anaphor (*myself, himself, themselves*) is bound in its governing category.
> **Principle B**: A pronominal (*me, he, him, them*) is free in its governing category.

In Chomsky's setting, 'bound' is defined in terms of c-command, and the 'governing category', the domain of binding, is roughly the clause. However, subsequent work has shown that other syntactic domains are also relevant for the definition of binding constraints, and that

the binding domain can vary for different pronominal elements, even within the same language.

2.1 *The binding domain*

We adopt the binding theory of Dalrymple (1993, 2001), who builds on original work by Bresnan *et al.* (1985) in proposing four domains which are relevant for anaphoric binding: the Root Domain, i.e. the domain consisting of an entire sentence or utterance; the Minimal Finite Domain, i.e. the minimal syntactic domain containing a finite element; the Minimal Complete Nucleus, i.e. the minimal syntactic domain containing an argument with the grammatical function subject; and the Coargument Domain, i.e. the minimal domain defined by a predicate and the arguments it governs. More recent work on binding theory in LFG adopts and builds on this approach (see in particular Bresnan *et al.* 2016, Chapters 9 and 10, and references cited there).

The binding domain for each anaphoric element is specified in its lexical entry. For example, the English personal pronouns obey binding constraints defined in terms of the Coargument Domain: a pronoun may not corefer with a superior coargument of the same predicate. In (3b), *him* may not be interpreted as coreferring either with *Alan* or *Bertie*, since all three are arguments of the same predicate, *told*.[2] In (3c,d), *him* may corefer with *Alan*, since they are arguments of different predicates: for example, in (3c), *Alan* is the subject of *saw*, and *him* is the object of *near*.

(3) a. Alan$_i$ likes him$_{*i}$.
b. Alan$_i$ told Bertie$_j$ about him$_{*i/*j}$.
c. Alan$_i$ saw a snake near him$_{i/j}$.
d. Alan$_i$ said that Bertie$_j$ likes him$_{i/*j/k}$.

In contrast, the antecedent of the English reflexive pronoun must corefer with a superior element within a different syntactic domain, the Minimal Complete Nucleus. In (4b), the reflexive pronoun rather than the personal pronoun is used to indicate coreference between *himself* and either *Alan* or *Bertie*; in (4c), coreference is allowed between

[2] Following standard practice, we use alphabetic subscripts such as i and j to indicate coreference. Later in the paper, we introduce indices as linguistic objects, and we represent the unique index of a phrase by a numeral.

himself and *Alan*, since both appear in the same Minimal Complete Nucleus; in (4d), coreference is disallowed between *Alan* and *himself*, since *Alan* appears outside the Minimal Complete Nucleus in which *himself* appears.

(4) a. Alan$_i$ likes himself$_i$.
b. Alan$_i$ told Bertie$_j$ about himself$_{i/j}$.
c. Alan$_i$ saw a snake near himself$_i$.
d. Alan$_i$ said that Bertie$_j$ likes himself$_{*i/j}$.

These examples illustrate the differing nature of the binding constraints on English personal and reflexive pronouns, encoded in Chomsky's Principles A and B as the difference between bound and free anaphoric elements. Anaphors such as *himself* obey *positive* constraints, requiring a particular syntactic relation to hold between anaphor and antecedent (i.e. that the antecedent must be bound by a superior element within the anaphor's binding domain). In contrast, pronominals such as *him* obey *negative* constraints, ruling out certain syntactic relations from holding between the pronominal and the superior elements within the relevant domain. As we will see in the following sections, positive constraints on a reflexive pronoun like *himself* are simpler to state than negative constraints on a personal pronoun like *him*, particularly when plural reference is brought into the picture.

2.2 *Superiority*

Besides specification of the binding domain, we must also specify the elements within the domain which are relevant for binding constraints: these are the elements which are *superior* to the anaphoric element within the domain. Superiority is defined in terms of both structural configuration and grammatical prominence.

Structurally, we take functional command to be the relevant configuration. We will return to the exact formalization of this relation in LFG in Section 7, but in theory-neutral terms, an element x functionally commands y iff x and y are coarguments or y is embedded in a coargument of x.

A grammatical prominence condition is also relevant: for example, although the subject and object of a transitive predicate functionally command each other, the subject is more grammatically promi-

nent than the object. Thus, an object reflexive may be bound by a subject coargument, but a subject reflexive may not be bound by an object coargument. For simplicity, we take the relevant prominence condition to be the grammatical function hierarchy (Keenan and Comrie 1977):[3] a subject binds its coarguments and elements contained in its coarguments, but an object does not bind the subject or elements contained in the subject.

3 NEGATIVE CONSTRAINTS AND COREFERENCE

Although it is certainly true that the negative constraint on a pronoun like *him* is stateable in terms of the syntactic domain in which it appears, it is vital to note that the constraint against identity of reference with a coargument cannot be enforced simply by constraining the choice of antecedent; for example, by disallowing an anaphor taking a coargument as antecedent. Consider example (5):

(5) Bertie thought that he had seen him.

Here, *he* and *him* are coarguments, and *he* is superior to *him*; therefore *him* may not take *he* as its antecedent. *Bertie* is not a coargument of either *he* or *him*, and so in principle *Bertie* may serve as antecedent for either pronoun. That is, *he* may take *Bertie* as antecedent, and likewise *him* may take *Bertie* as antecedent. However, as observed by Wasow (1972), Higginbotham (1983), and Lasnik (1989c), *Bertie* may not function as antecedent to *he* and *him* simultaneously, since this would result in coreference between *he* and *him*, and this is not allowed. Note that such a configuration is not ruled out by simple syntactic constraints on where the antecedent of each pronoun can appear. Thus, although the basic constraint is syntactic, its application requires a semantic/pragmatic resolution of reference: the individuals referred to by *he* and *him* in (5) may not be the same. Our analysis improves on previous work in LFG in explicitly defining the appropriate notion of coreference, and using this definition in the statement of negative constraints.

[3] See Dalrymple (1993, Chapter 5) and Bresnan *et al.* (2016, 218, 246–247, 276) for discussion of additional conditions that have been shown to be relevant to defining grammatical prominence, including the role of linear precedence relations and the thematic hierarchy.

4 PLURALITY AND BINDING REQUIREMENTS

Plural reflexives are subject to the same positive constraint as singular reflexives: a plural reflexive must corefer with a superior antecedent within the Minimal Complete Nucleus, and long-distance or split antecedents are not acceptable.

(6) a. [Alan$_i$ and Bertie$_j$]$_{i+j}$ like themselves$_{i+j}$.
b. *[Alan$_i$ and Bertie$_j$]$_{i+j}$ said that Charlie$_k$ likes themselves$_{i+j}$.
c. *Alan$_i$ confronted Bertie$_j$ with themselves$_{i+j}$.

In (6), the proper names *Alan* and *Bertie* each bear an index, and the coordinated phrase *Alan and Bertie* also bears a separate, complex index constructed from the conjuncts, as we discuss in Section 9.1. The reflexive must have the same (simple or complex) index as its antecedent.

The situation with plural personal pronouns is considerably more complex. Like singular personal pronouns, plural personal pronouns obey a negative constraint: in (7), *them* may not have the same index as its coargument *Alan and Bertie*, just as in example (3).

(7) *[Alan$_i$ and Bertie$_j$]$_{i+j}$ like them$_{i+j}$.

A nonoverlapping (disjoint) relation between the pronoun and its coarguments is uncontroversially acceptable, similar to the requirement for singular pronouns to be noncoreferent with coarguments, as shown in (8).

(8) [Alan$_i$ and Bertie$_j$]$_{i+j}$ like them$_{k+l}$.

However, with plural pronouns and plural coarguments, other patterns are possible:

A. The index of the coargument is properly included in the index of the pronoun:
[Alan$_i$ (and Bertie$_j$)]$_{i/i+j}$ like(s) them$_{i+j+k}$.
B. The index of the pronoun is properly included in the index of its coargument:
[Alan$_i$, Bertie$_j$, and Charlie$_k$]$_{i+j+k}$ like him$_i$/them$_{i+j}$.
C. The index of the pronoun overlaps with the index of the coargument, but without an inclusion relation:
[Alan$_i$ and Bertie$_j$]$_{i+j}$ like them$_{j+k}$.

A fourth pattern has been claimed to be relevant in some of the literature on pronominal binding:

D. The index of the pronoun is the sum of the indices of the coarguments, but not identical to any coargument:
Alan$_i$ told Bertie$_j$ about them$_{i+j}$.

The grammatical status of these patterns is controversial, and various positions have been taken in the literature as to their acceptability, as we now outline. To avoid confusion due to the varying judgements that have been reported, we explicitly mark each example with the judgement reported in the cited work.

4.1 *Pattern A: Index of coargument is properly included in index of pronoun*

Some simple examples conforming to this pattern have been judged as ungrammatical:

(9) a. *He$_i$ represented them$_{i+}$. (Seeley 1993, 309)
b. *Bill$_i$ represented them$_{i+}$. (Seeley 1993, 309)
c. *John$_i$ told them$_{i+j}$ that Mary$_j$ should leave. (Lasnik 1989a, 151)

However, many other examples conforming to this pattern have been judged as acceptable:

(10) a. ✓ He$_i$ talked about them$_{i+}$. (Fiengo and May 1994, 43)
b. ✓ Bill$_i$ was quite pleased [that Mary$_j$ defended them$_{i+j}$]. (Seeley 1993, 308)
c. ✓ Bill$_i$ was happy [because Mary$_j$ had protected them$_{i+j}$]. (Seeley 1993, 308)
d. ✓ Sam$_i$ is telling Tom$_j$ not to praise them$_{i+j}$. (Seeley 1993, 308)
e. ✓ Bill$_i$ was surprised that [Mary$_j$'s representing them$_{i+j}$ at the trial] had caused such problems. (Seeley 1993, 308)
f. ✓ John$_i$ wants Mary$_j$ to represent them$_{i+j}$. (Berman and Hestvik 1997, 5)
g. ✓ [John$_i$'s mother]$_j$ protected them$_{i+j}$ from the robbers. (Berman and Hestvik 1997, 6)
h. ✓ [The woman who loved John$_i$]$_j$ represented them$_{i+j}$ at the trial. (Berman and Hestvik 1997, 6)

i. ✓ [Bill$_i$ and Mary$_j$]$_{i+j}$ were asked to appear before the committee. But Bill$_i$ fell ill and had to be excused. John$_k$ said that Mary$_j$ represented them$_{i+j}$. (Berman and Hestvik 1997, 7)

It has been claimed that increasing the number of coarguments whose reference is included in the reference of the pronoun degrades acceptability; Seeley (1993) provides the following judgements:

(11) a. ?The doctor$_i$ told the patient$_j$ [that the nurse$_k$ would protect them$_{i+j+k}$ during the storm]. (Seeley 1993, 313)

b. ??The doctor$_i$ said [that the patient$_j$ told the nurse$_k$ about them$_{i+j+k}$]. (Seeley 1993, 313)

Nevertheless, the large number of acceptable examples of Pattern A indicate that syntactic constraints do not rule out this pattern, but that unacceptable examples are ruled out by some combination of semantic or pragmatic conditions.

Several instances of this pattern reported in the literature involve first-person singular subjects with a coargument first-person plural personal pronoun:

(12) a. ✓ I expect us to meet John at the party. (Fiengo and May 1994, 44)

b. ✓ I believe us to have been cheated. (Kiparsky 2002, 20)

c. ✓ I prefer to call us rape statistics. (Kiparsky 2002, 20)

d. ✓ I want us to be friends. (Kiparsky 2002, 21)

e. ✓ We have a terrific team. I really like us. (Kiparsky 2002, 19)

For reasons that are not clear to us, reported judgements of the reverse pattern (*We…me*) are more often judged as unacceptable, as we discuss below.

4.2 *Pattern B: Index of pronoun is properly included in index of coargument*

Some examples of Pattern B have also been judged as ungrammatical:

(13) a. *They$_{i+}$ like him$_i$. (Lasnik 1989b, 125; Seeley 1993, 309)

b. *[John$_i$ and Mary$_j$]$_{i+j}$ are taking care of him$_i$. (Kiparsky 2002, 20)

However, just as with Pattern A, many examples of this pattern have been judged as acceptable:

(14) a. ✓ [Richard$_i$ and Pat$_j$]$_{i+j}$ both regard him$_i$/her$_j$ as innocent. (Kiparsky 2002, 20)

b. ✓ [John$_i$ and Mary$_j$]$_{i+j}$ talked about him$_i$. (Fiengo and May 1994, 43)

c. ✓ [John$_i$ and Mary$_j$]$_{i+j}$ often connive behind their colleagues' backs to advance the position of one or the other. This time, they$_{i+j}$ managed to get her$_j$ a position in the front office. (Berman and Hestvik 1997, 8)

d. ✓ [The men$_i$ and Mary$_j$]$_{i+j}$ talked about them$_i$. (Fiengo and May 1994, 43)

e. ✓ John$_i$ and Mary$_j$ discussed their participation in the upcoming contest with Bill$_k$. They$_{i+j+k}$ expect them$_{i+j}$ to win. (Berman and Hestvik 1997, 8)

f. Acceptable "for many speakers": [Felix$_i$ and Lucie$_j$]$_{i+j}$ authorized her$_j$ to be their representative. (Reinhart and Reuland 1993, 677)

Distributivity has been claimed to be a factor in the acceptability of Pattern B examples. Reinhart and Reuland (1993) claim that there is a contrast between the examples in (15), where *both* forces a distributive reading:

(15) a. ✓ [Max$_i$ and Lucie$_j$] talked about him$_i$. (Reinhart and Reuland 1993, 677)

b. *[Both Max$_i$ and Lucie$_j$] talked about him$_i$. (Reinhart and Reuland 1993, 677)

According to Reinhart and Reuland (1993) and Kiparsky (2002), Pattern B sentences are ungrammatical only under a distributive reading, but are fine under a collective reading. Strikingly, however, Seeley (1993) judges the examples with *both* in (16b,c) as grammatical, in contrast to (16a), which he judges as ungrammatical. Seeley observes that the presence of *both* plays a 'key role'.

(16) a. *They$_{i+}$ like him$_i$. (Seeley 1993, 309)

b. ✓ [Bill$_i$ and Mary$_j$]$_{i+j}$ both introduced him$_i$. (Seeley 1993, 308)

c. ✓ They$_{i+}$ both introduced him$_i$. (Seeley 1993, 308)

Again, then, we take the large number of acceptable Pattern B examples, together with the lack of agreement about distributivity, as demonstrating that Pattern B examples do not violate syntactic constraints, but should be accounted for in semantic or pragmatic terms.

As noted above, first-person examples seem to differ to some degree between Pattern A and Pattern B, in that more Pattern B examples than Pattern A examples are judged as ungrammatical.

(17) a. *We like me. (Lasnik 1989b, 125)
b. *We watched me leaving (in the mirror). (Chomsky 1973, cited in Lasnik 1989d)

Fiengo and May (1994) judge (18b) as 'worse' than (18a), its Pattern A counterpart:

(18) a. ✓ I expect us to meet John at the party. (Fiengo and May 1994, 44)
b. 'worse': We expect me to meet John at the party. (Fiengo and May 1994, 44)

Nevertheless, some Pattern B first-person examples are judged as acceptable:

(19) a. ✓ We made John president and me vice-president. (Fiengo and May 1994, 44)
b. ✓ By an overwhelming majority, we preferred me. (Kiparsky 2002, 19)

4.3 *Pattern C: Index of pronoun overlaps with index of coargument without inclusion*

Although much attention has been paid to cases A and B where the index of the pronoun properly includes the index of a coargument or vice versa, no one to our knowledge has discussed the third logically possible pattern of overlapping indices, where neither index properly includes the other. We therefore find no such examples in the literature, but we can easily construct them with appropriate changes to the examples in sections A and B:

(20) a. Bill$_i$ was happy [because [Mary$_j$ and her dog$_k$]$_{j+k}$ had protected them$_{i+j}$]. (≈ 10c)

b. [Richard$_i$ and Pat$_j$]$_{i+j}$ both regard [him$_i$ and his wife$_k$]$_{i+k}$ as innocent. (≈ 14a)

We judge these examples acceptable. And to the extent that the original examples (10c) and (14a) have been judged acceptable, there is no reason to expect a different judgement in this case.

4.4 *Pattern D: Index of pronoun is sum of indices of coarguments*

This pattern is rarely discussed, though it is a focus of attention in work by Seeley (1993) and Berman and Hestvik (1997). Note that Pattern D is only possible with predicates that can take more than two arguments, since the pronoun must fill one argument slot and refer to a group composed of individuals that fill at least two of the other argument slots. This more or less excludes one common class of ditransitives, the transfer verbs, as a source of examples, because the object is typically inanimate and the recipient/goal typically animate. Fiengo and May (1994) provide the example in (21a) in support of their claim that Pattern D is acceptable, while Seeley provides the example in (21b) in support of the opposite claim. Berman and Hestvik (1997) discuss both examples and agree with both judgements, and also provide the example in (21c). They furthermore claim – without examples – that similar ungrammatical binding patterns can be constructed with the verbs *assign to, deny to, cede to, compare to, consign to, entrust to, explain to, leave to, offer to, point out to, promise to*, and *reveal to*. Finally, Kiparsky (2002) provides example (21d) with the judgement as indicated.

(21) a. ✓ John$_i$ talked to Mary$_j$ about them$_{i+j}$. (Fiengo and May 1994, 40; Berman and Hestvik 1997, 24)

b. *Bill$_i$ told Mary$_j$ about them$_{i+j}$. (Seeley 1993, 307; Berman and Hestvik 1997, 6)

c. (At their wedding reception, John and Mary were speaking to Bill and Sue.) *John$_i$ said that he$_i$ wanted [PRO$_i$ to photograph Mary$_j$ for them$_{i+j}$]. (Berman and Hestvik 1997, 25)

d. *John$_i$ confronted Bill$_j$ with them$_{i+j}$. (Kiparsky 2002, 21)

In these examples, *them* does not corefer with either of its coarguments, but overlaps in reference with both such that its reference is

exhausted by its coarguments. Seeley (1993) and Berman and Hestvik (1997, 20–21, 24–27) claim that this is disallowed, and that it is necessary not only to prohibit coreference between the discourse referents introduced by a pronoun and a (superior) coargument, but also between the pronoun's discourse referent and the sum of its (superior) coarguments. Berman and Hestvik also discuss (21a) and judge it as acceptable, claiming that its acceptability is due to special properties of *talk* that remain unclear.[4]

Nevertheless, the evidence seems inconclusive to us. Berman and Hestvik (1997, Section 5) admit that the empirical status of these examples is somewhat unclear. For what it is worth, note that it is possible to find naturally occurring examples with first and second person arguments as in (22).

(22) Khushi looks up at Arnav with tear-filled eyes: *Would you tell me about us?* How did we meet? When did we fell in love? Everything from the beginning?
(http://fast-forward-by-tia.blogspot.no/)

The theory that we develop below predicts that Pattern D is grammatical, although it is possible to rule it out in our model; we return to this issue in Section 9.5.

In sum, previous work on binding involving plural pronouns or antecedents is unanimous in ruling out strict coreference between a plural pronoun and a superior coargument, but there is a great deal of variation in judgements on cases of overlap or inclusion (for additional discussion of this point, see Büring 2005, Chapter 9). It is well known that binding possibilities are influenced by lexical, structural, and contextual factors that are not yet completely understood; see Jackendoff (1992), Reinhart and Reuland (1993), Berman and Hestvik (1997), and Park (2012) for discussion. We take the position that syntactic binding constraints for English singular and plural personal pronouns rule out coreference between the pronoun and its superior coarguments, but that overlap or inclusion is permitted. This means that the unacceptable examples in this section that do not involve strict coref-

[4] They do, however, rule out an analysis according to which *Mary* does not c-command out of the PP in (21a) because they judge *John talked to Mary$_i$ about herself$_i$* acceptable.

erence with a coargument are not ruled out by syntactic constraints, but are unacceptable for other reasons.

5 FORMALISING ANAPHORIC RESOLUTION

The most successful attempts to deal with anaphoric resolution, especially across sentences, have been developed within the tradition of dynamic semantics. We follow that tradition here, in particular the version developed by Kamp and Reyle (1993) and Kamp *et al.* (2011), Discourse Representation Theory (DRT).[5] In dynamic semantics, the meaning of a sentence is not its truth conditions, but its *context change potential*, made precise as a relation between assignments of individuals to discourse referents at different points in the discourse. Consider the DRS in (23).

(23) A linguist arrived.

x
$\textit{linguist}(x)$ $\textit{arrive}(x)$

(23) is interpreted as a relation between an 'input' assignment i and an 'output' assignment o such that o is like i except it assigns some individual to x that is in the denotation of *linguist* and *arrive*.[6] This is shown in (24), where $\mathfrak{I}$ is the interpretation function assigning relational meanings to predicate constants and $i \subset_{\{x\}} o$ means that o is like i except in assigning some value to x.

(24) Interpretation of (23) as a relation between input and output assignments (Kamp and Reyle 1993):

$$\{\langle i, o\rangle | i \subset_{\{x\}} o \wedge o(x) \in \mathfrak{I}(\textit{linguist}) \wedge o(x) \in \mathfrak{I}(\textit{arrive})\}$$

Although in this setting the meaning of a sentence is a relation between assignments, there is a natural way to get to truth conditions,

[5] Similar ideas are found in Heim (1982) and many later versions of dynamic semantics.

[6] We use i and o as variables over states when these function as input and output states of a DRS, but s when we talk about states more generally.

by taking i to be the empty assignment and requiring the existence of some assignment o. That is, we define truth (using $s_\emptyset$ for the empty assignment) so that (23) is true iff there is an o such that $\langle s_\emptyset, o\rangle$ is in (24). It is easily seen that this yields the same truth conditions as for the first-order translation of (23), which is $\exists x.\mathit{linguist}(x) \wedge \mathit{arrive}(x)$.

Nevertheless, although the truth conditions turn out the same as in first-order logic, there is a difference in the predictions about anaphoric accessibility. The idea is to use o, the 'output' assignment of this sentence, as the input assignment for the subsequent discourse, thereby making x accessible for anaphoric uptake.

(25) He sat down.

y
$\mathit{sit.down}(y)$ $y = ?$

Here the anaphor is associated with a condition $y = ?$, which we can interpret as an instruction to find an antecedent. If this sentence follows sentence (23), y can be equated with x, with the result that the two-sentence discourse means that there is some individual who is a linguist, arrived, and sat down.

However, this treatment of anaphora means that the DRS in (25) has no interpretation at all until such an antecedent is found. This makes the framework representational: the DRSs are essential ingredients of the analysis and cannot be 'translated away' the way lambda terms can be in the Montagovian tradition. Moreover, because (25) has no meaning until the antecedent is found, we cannot make sense of the intuition that, in many cases, the sentence containing the anaphor will constrain the resolution. For example, it is likely that *it* in *It meowed* will be resolved to some animal making the appropriate sound, but we cannot model this if *It meowed* does not have a meaning until a referent is found.

This representational nature of Kamp and Reyle's DRSs was an obstacle to compositionality. In response, Muskens (1996) developed a compositional version of DRT, CDRT. The leading idea is to inject assignments into the object language, with explicit quantification over information states, plus an 'interpretation function' v which assigns an

individual inhabitant to each discourse referent in every state. That is, discourse referents (or registers, as they are often called in CDRT) are no longer simply variables over individuals but are reified as terms of a separate type, π, which are 'interpreted' by the function ν. Furthermore, we have axioms which guarantee that ν actually works as an assignment.

With that in place, we can view DRSs such as (23) as abbreviations for more complex lambda terms. Instead of interpreting the DRS as a relation between assignments in the metalanguage, we now expand it as $\lambda i.\lambda o.P$ in the object language, where P is the contents of the DRS. As before, those contents have two parts: a universe and a set of conditions. In the conditions, we expand a discourse referent x as $\nu(o)(x)$. Observe that x here is a constant (of type π), but when we plug it into the ν function, we get a term of type e, the inhabitant of that discourse referent. So in practical terms x works like a variable. We interpret the declaration of discourse referents in the universe of a DRS as a constraint that the input and output states of that DRS, i and o, differ at most with respect to the values of those variables, i.e. (for (23)), $\forall\delta.\delta \neq x \rightarrow \nu(i)(\delta) = \nu(o)(\delta)$.

In sum, we now have (26) as the expansion of (23).

(26) Content of (23) in CDRT (Muskens 1996):

$$\lambda i.\lambda o.\forall\delta.\delta \neq x \rightarrow \nu(i)(\delta) = \nu(o)(\delta) \wedge \textit{linguist}(\nu(o)(x)) \wedge \textit{arrive}(\nu(o)(x))$$

Compared to (23)–(24), what has happened here is that the assignments, which only played a role in the metalanguage interpretation (24) in the DRT approach, are now part and parcel of the object language (26). Nevertheless, we can get to the truth conditions in a very similar way by saturating i with the empty assignment $s_\emptyset$ and existentially closing o.

What about the unresolved anaphor in (25)? It is not trivial to give a model-theoretic semantics for a condition like $y = ?$, which seems irreducibly procedural: first we pick an antecedent and then we interpret the whole thing semantically. Muskens' solution was to simply use coindexation, which in CDRT terms means that we use the same discourse referent for both *he* and *a linguist*. But this means the syntax has to take care of anaphoric resolution, which is problematic for several reasons, as noted by Beaver (2002).

We therefore follow Haug (2014b), who partialized the underlying logic to allow for model-theoretic representation of unresolved anaphora in his Partial CDRT (PCDRT). In a given state, ν now acts as a *partial* assignment, which means that we can identify unused discourse referents. Instead of using a constant x, CDRT uses a function expression picking out the first unused discourse referent in the input state (i). We forego details here, but the reader should bear in mind that x in (26) is not in fact a free variable (or a constant), but a discourse referent functionally dependent on i.

More importantly, anaphoric discourse referents are translated as any other, without any coindexation. They are, however, marked as anaphoric; we represent anaphoric discourse referents with an overbar ($\bar{x}$). The truth definition then requires all anaphoric discourse referents to corefer with an accessible antecedent, as in (27): otherwise there is a truth value gap. This latter effect is achieved by Beaver's unary presupposition connective ∂ (Beaver 1992), which maps $\partial(\phi)$ to true if ϕ is true and to the undefined truth value otherwise.

(27) Condition on antecedency for anaphoric discourse referents:
$\partial(\nu(s)(\bar{x}) = \nu(s)(\mathscr{A}_s(\bar{x})))$

This condition requires $\bar{x}$ to be identical to its antecedent $\mathscr{A}_s(\bar{x})$ in the state s, as specified by the antecedency function $\mathscr{A}$, thus yielding coreference or, if $\mathscr{A}_s(\bar{x})$ is itself bound by an operator, covariation. Notice that $\bar{x}$ and its antecedent must both be defined in the same state s; this yields the usual operator-induced restrictions on anaphoric accessibility, as in DRT. We often omit the subscript s on the anaphoric relation $\mathscr{A}$, while retaining the requirement that the anaphoric relation $\mathscr{A}$ is defined only between discourse referents in the same state.[7]

$\mathscr{A}$ is a function from discourse referents in a particular state s to discourse referents (in the same state s). It is a composite function:

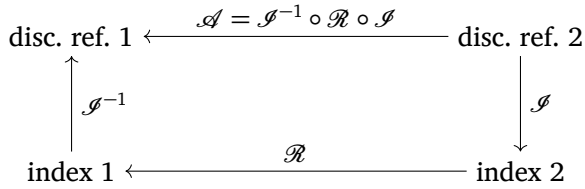
(28) Definition of $\mathscr{A}$ in a state s:[8]
$\mathscr{A}_s(x) \equiv \mathscr{I}_s^{-1}(\mathscr{R}(\mathscr{I}_s(x)))$

[7] Haug (2014b) in addition assumes a constraint $\mathscr{A}_s(x) < x$ requiring the antecedent to precede the pronoun, but the (marked) possibility of cataphoric resolution shows that this constraint is non-monotonic.

[8] Haug (2014b, 497, ex. 69) defines $\mathscr{A}$ in terms of $\mathscr{R}^*$, the transitive closure of $\mathscr{R}$. We define and discuss $\mathscr{R}^*$ in Section 7.

The function $\mathscr{I}_s$ maps the discourse referents in a state s[9] to objects which we will call 'indices', which introduce discourse referents; crucially, as we will see, indices are accessible to syntactic representations and constraints. $\mathscr{I}^{-1}$ is the inverse mapping, a function from indices back to the discourse referents they introduce (in a particular state s). The core of pragmatic anaphora resolution is then the function $\mathscr{R}$, which maps indices to antecedent indices. This allows us to keep the simple idea underlying the coindexation approach, namely that anaphoric relations are just relations between syntactic tokens, but without presupposing that the resolution is actually done in the syntax.[10] We thus have the following set-up: indices, which are syntactically accessible, introduce discourse referents; by mapping from discourse referents to indices, then from indices to antecedent indices, and finally from antecedent indices to discourse referents, we obtain a mapping between discourse referents and their antecedent discourse referents (in a particular state).

(29) The relations $\mathscr{A}$, $\mathscr{R}$ and $\mathscr{I}$:



Since $\mathscr{A}$ is uniquely determined by $\mathscr{R}$, we will use constraints on $\mathscr{R}$ to capture the constraints of binding theory. But first, in the following section, we integrate the model with the framework of LFG.

6 INTEGRATING SYNTAX

In this section we show how the PCDRT approach to anaphora can be integrated with the grammatical framework of Lexical Functional Grammar (LFG), to provide a formal model of the interaction between

[9] As with $\mathscr{A}$ we will often omit the subscript s on $\mathscr{I}$.

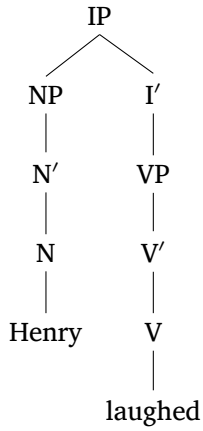
[10] There are also technical advantages over the view that anaphoric relations hold between discourse referents in context, because the semantics quantifies over contexts in a way that would scope over anaphoric resolutions, whereas anaphoric resolution between indices/syntactic tokens will always be scopeless. See Nouwen (2003, 140) and Haug (2014b, 482–483) for details.

syntax and pragmatics. LFG is a constraint-based, non-derivational framework for grammatical analysis, developed by Kaplan and Bresnan (1982), and presented in detail by e.g. Dalrymple (2001), Falk (2001) and Bresnan *et al.* (2016). A crucial element of the LFG framework is that different types of grammatical information are distinguished from one another and treated as distinct levels of grammatical representation, related by means of piecewise functions called *projections*. LFG therefore provides an ideal grammatical framework into which to integrate PCDRT, with its clear representational separation of semantics and pragmatics.

For example, the phrasal structure of a clause, the *c(onstituent)-structure*, is treated as one level of grammatical representation, represented by means of a phrase-structure tree. In contrast, functional syntactic relations, e.g. grammatical functions such as subject and object, are treated at a separate level, *f(unctional)-structure*, represented as an attribute-value matrix. So, for the English sentence in (30), the surface phrasal structure, the c-structure, can be represented as in (31), and the abstract syntactic structure, the f-structure, can be represented as in (32). Following standard LFG conventions, we represent only those features of f-structure that are relevant for the discussion at hand, omitting features encoding information about person, number, gender, tense, aspect, and other grammatical information.

(30) Henry laughed.

(31)



(32) $\begin{bmatrix} \textsc{pred} & \text{'laugh}\langle\textsc{subj}\rangle\text{'} \\ \textsc{subj} & \begin{bmatrix} \textsc{pred} & \text{'Henry'} \end{bmatrix} \end{bmatrix}$

These two grammatical modules are related via a projection function ϕ, which maps c-structure nodes to their corresponding f-structures. Constraints on the ϕ function are stated in terms of functional descriptions associated with nodes of the phrase structure tree; these functional descriptions use the variable $*$ to represent the c-structure node on which the constraint appears, and the variable $\hat{*}$ to represent the mother of the node bearing the constraint. The f-structure projected from a c-structure node is therefore obtained by applying the function ϕ to $*$, i.e. $\phi(*)$, and the f-structure projected from a c-structure node's mother is obtained by applying ϕ to $\hat{*}$, i.e. $\phi(\hat{*})$. These functions are usually abbreviated by the f-structure metavariables $\downarrow$ and $\uparrow$:

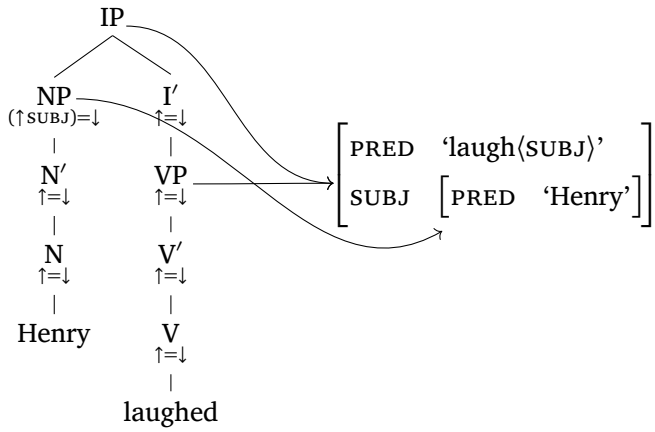
(33) a. $\downarrow \equiv \phi(*)$
b. $\uparrow \equiv \phi(\hat{*})$

These metavariables enable concise statements of the constraints on the relation between c-structures and f-structures. For example, in English the specifier of IP is associated with the grammatical role of subject. We represent this by means of the following phrase-structure rule:

(34) $\begin{array}{lcccc} \text{IP} & \rightarrow & \text{NP} & & \text{I}' \\ & & (\uparrow \textsc{subj}) = \downarrow & & \uparrow = \downarrow \end{array}$

The annotation $(\uparrow \textsc{subj}) = \downarrow$ on the constituent in the specifier of IP requires that the f-structure projected from the NP ($\downarrow$) supply the value of the attribute SUBJ in the f-structure projected from the NP's mother ($\uparrow$). The annotation $\uparrow = \downarrow$ on the I′ requires the f-structure projected from the I′ ($\downarrow$) and the f-structure projected from the IP ($\uparrow$) to be the same. Ex. (35) shows the c-structure for (30), just as in (31) but with annotated constraints referring to the f-structure on each node. The f-structure is as in (32), and the projection function ϕ is represented by means of arrows between the two structures.

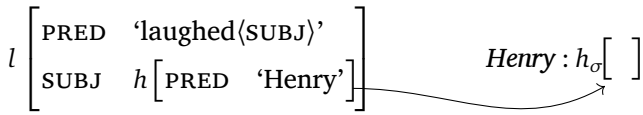
(35)



The concept of projection functions between different levels of grammatical representation was generalised by Kaplan (1989) in terms of a 'projection architecture' modelling the different levels of linguistic structure and the relations among them. The full inventory of levels of grammatical representation and the projection functions relating them are a matter of debate, but the details do not concern us here.[11]

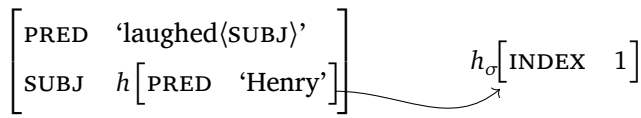
Of crucial importance for the present topic, however, is the interface between syntax and semantics. Work on semantics in LFG makes use of the 'glue' theory of the syntax-semantics interface (Dalrymple 2001; Asudeh 2012), according to which meanings are paired with logical expressions which constrain their composition. In standard approaches to glue semantics within LFG, meanings are paired with logical formulae over *s(emantic)-structures*, projected from f-structures via the projection function σ. For example, the meaning of the proper name, *Henry*, is paired with a semantic structure projected from the SUBJ f-structure. For ease of exposition, we introduce labels such as l and h to facilitate reference to different parts of the f-structure. As is standard, we use a subscript σ to refer to the s-structure projected from a given f-structure. Thus, h_σ is the semantic structure corresponding to the f-structure labeled h.

[11] On the projection architecture of LFG, see e.g. Bögel *et al.* (2009), Dalrymple and Mycock (2011), Dalrymple and Nikolaeva (2011), Giorgolo and Asudeh (2011), Asudeh (2012, 53), and Mycock and Lowe (2013).

(36) 

S-structure is an interface structure for modelling the influence of syntax on semantic compositionality. Recent work on semantic structure has emphasised its internal complexity, particularly in regard to the embedding of s-structures within other s-structures (Asudeh and Giorgolo 2012; Asudeh *et al.* 2014) and the types of features that are present within s-structures (Dalrymple and Nikolaeva 2006).

We propose that the indices between which anaphoric relations hold in PCDRT are a component of semantic structure. It has been observed in the glue literature (e.g. Kokkonidis 2008, 63) that the empty 'placeholder' semantic structures typically used in (higher-order) glue semantics would – under the standard, set-theoretic interpretation of LFG attribute-value structures – in fact lead to an unwanted lack of differentiation among semantic structures. To guarantee that we can keep semantic structures apart, it is necessary to equip them with a uniquely identifying element working in much the same way as the semantic form value of the PRED feature at f-structure (Kaplan and Bresnan 1982, 225). We take these uniquely identifying elements to be the indices discussed in the previous section, and we assign indices as the values of the s-structure feature INDEX.

(37) 

In (37), then, the index 1 uniquely identifies the semantic structure h_σ. We assume that all semantic structures that introduce discourse referents are associated with an index (though, as we will see, not all indices are associated with semantic structures: complex indices can also be constructed by combining contextually salient indices). Indices might also be associated with semantic structures introducing discourse referents over events or times, but for the purposes of binding theory we are only interested in type e (nominal) discourse referents.

Thus, the value of (h_σ INDEX) is a unique index that is mapped to some discourse referent in a given information state s by the func-

tion $\mathscr{I}_s^{-1}$ discussed above. In complex contexts involving embedded DRSs and hence several information states, there will be several functions $\mathscr{I}_s^{-1}$ potentially mapping different semantic indices to the same discourse referent (interpreted in different states): see Haug (2014b) for more details on how this works. For our purposes, however, we do not need to deal with embedded DRSs or different information states, and we can therefore make the simplifying assumption that semantic indices map one-to-one to discourse referents. To ease the exposition we can use integers n for the values of INDEX attributes, and x_n for the corresponding discourse referents.

In the next sections we show how syntactic constraints on the interpretation of pronouns can be defined in terms of an $\mathscr{R}$ relation between indices.

7 REFLEXIVE PRONOUNS AND POSITIVE CONSTRAINTS

We begin with a relatively simple example of the positive binding constraint on English reflexive pronouns, before moving on to consider the more complex issue of negative constraints. In this section, it will largely be sufficient to adapt existing machinery and analyses to our setting. This will introduce standard aspects of LFG's binding theory, which we then extend to deal with negative constraints.

The positive binding constraint on English reflexives is stated in (38):

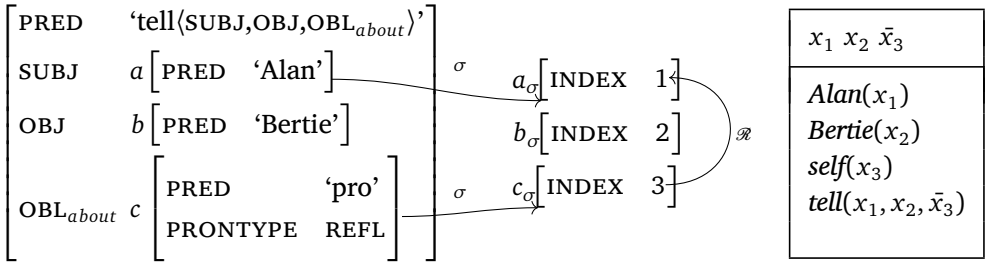
(38) English reflexives must have a superior antecedent within the Minimal Complete Nucleus binding domain.

The Minimal Complete Nucleus is the minimal f-structure containing a SUBJ function. This means that in the following example, the reflexive pronoun *himself* may in principle corefer with either the subject *Alan* or the object *Bertie*.

(39) Alan$_i$ told Bertie$_j$ about himself$_{i/j}$.

Let us assume that in context the most natural interpretation is where *himself* is coreferent with the subject, *Alan*. As discussed in the previous section, the indices of words introducing discourse referents appear as the value of the feature INDEX in the s-structure of the word concerned. In this case, then, we have the following $\mathscr{R}$ relation:

(40) The $\mathscr{R}$ relation: Alan$_i$ told Bertie about himself$_i$.



The index 1 maps to the discourse referent x_1 by the inverse function $\mathscr{I}_s^{-1}$, 2 maps to x_2, and 3 maps to x_3. In the context under consideration, the index 1 of the subject *Alan* is assigned as the antecedent of the index 3 belonging to the reflexive pronoun; this is modelled by means of the function $\mathscr{R}$. In another context, the resolution $\mathscr{R}(3) = 2$ (the reflexive is bound by the object) could have been more likely. However, the grammar rules out the possibility that the reflexive has e.g. a sentence-external antecedent, with $\mathscr{R}(3)$ resolved to an index other than 1 or 2. So we want to recast (38) as a constraint on the resolution of $\mathscr{R}$.

In order to state (38) as a constraint on $\mathscr{R}$ in the LFG formalism, we need to express the notions of *superiority* and *binding domain*. The latter concept is relatively straightforward as it can be expressed by a formula of the general form shown in (41):[12]

(41) General relation between an anaphor with f-structure ↑ and its binders:
((GF^+ ↑) $\mathrm{GF}_{\mathrm{ant}}$)

This expression represents the set of potential f-structure antecedents of the reflexive pronoun *himself*. In this expression, ↑ is the f-structure corresponding to the preterminal node dominating the word *himself*; GF is a variable over grammatical functions (SUBJ, OBJ, etc.); GF^+ is a sequence of grammatical functions GF, a path through the f-structure ending in ↑; and $\mathrm{GF}_{\mathrm{ant}}$ is the grammatical function of the antecedent. The expression (GF^+ ↑) refers to any f-structure properly containing

[12] For a full explication, see Dalrymple (1993, 2001) and Bresnan *et al.* (2016).

the f-structure ↑, and the antecedent of the reflexive bears the grammatical function GF_{ant} within the f-structure (GF^{+} ↑).

(42) Schematic syntactic relation between the anaphor and its antecedent:

$$f \begin{bmatrix} \text{GF}_{\text{ant}} & [\text{ANTECEDENT}] \\ \text{GF} & \ldots\ [\text{REFLEXIVE } (\uparrow \text{ in } (41))] \end{bmatrix}$$

Notice that the form of this constraint makes sure that the antecedent functionally commands the reflexive as defined in Section 2.2: the antecedent bears the grammatical function GF_{ant} within some f-structure f, and the reflexive is embedded within f to some depth defined by the path GF^{+}. This means that the reflexive and its antecedent are either coarguments within f, or the reflexive is embedded inside a coargument of the antecedent (if GF^{+} has more than one element).

In order to impose the requirement for the reflexive to be bound within its binding domain, we must place the appropriate constraints on the path GF^{+} in (41). The English reflexive *himself* must be bound within the minimal complete nucleus (the minimal f-structure with a SUBJ function). This requirement is imposed by defining the path as MCNPATH:

(43) Minimal Complete Nucleus binding domain:

$$\text{MCNPATH} \equiv \begin{matrix} \text{GF}^{*} & \text{GF} \\ \neg(\rightarrow \text{SUBJ}) & \end{matrix}$$

The definition of MCNPATH in (43) contains an *off-path constraint*, ¬(→ SUBJ): off-path constraints appear as annotations on an attribute, and allow reference to the f-structure value of the attribute (→) or to the f-structure in which the attribute appears (←). The off-path constraint ¬(→ SUBJ) is interpreted as constraining each non-final grammatical function GF on the path, ensuring that MCNPATH does not pass through an f-structure with a SUBJ attribute. Other binding domains involve different off-path constraints on GF^{+}, as we will see.

Besides limiting the domain to the minimal f-structure containing a SUBJ, we must also make sure that GF_{ant} is constrained to range over grammatically more prominent elements within the binding domain. This prominence condition is imposed by ensuring that the f-structure value of GF_{ant} (the antecedent of the reflexive, labeled a in

the schematic diagram in (44)) is superior on the grammatical function hierarchy to its coargument GF which contains the reflexive, labeled c in (44).

(44) $$\begin{bmatrix} \mathrm{GF}_{\mathrm{ant}} & a\ [\textsc{antecedent}] \\ \mathrm{GF} & c\left[\ldots\ [\textsc{reflexive}]\right] \end{bmatrix}$$

The prominence condition has never been made explicit in the LFG literature. To state it, we first define the relation SUPERIOR, which holds between arguments of the same predicate:

(45) Definition of SUPERIOR:

SUPERIOR(f_1, f_2) if and only if f_1 and f_2 are arguments of the same predicate and f_1 outranks f_2 on the grammatical function hierarchy.

SUPERIOR constrains the relation between the f-structures labeled a and c in (44), requiring a to be superior to c. We can now impose the appropriate prominence condition by means of the following off-path constraints on $\mathrm{GF}_{\mathrm{ant}}$:

(46) Off-path constraints on $\mathrm{GF}_{\mathrm{ant}}$ encoding the superiority condition:

$$\begin{array}{c} \mathrm{GF}_{\mathrm{ant}} \\ \%\textsc{coarg} = (\leftarrow \mathrm{GF}) \\ (\%\textsc{coarg}\ \mathrm{GF}^{*}) = \uparrow \\ \textsc{superior}(\rightarrow, \%\textsc{coarg}) \end{array}$$

The constraints in (46) make use of a local name %COARG to refer to a coargument of the antecedent; local names are prefixed with a percent sign '%', and are used in order to ensure reference to the same coargument f-structure in each constraint. According to the first line, then, %COARG is defined as an f-structure bearing some grammatical function GF within the f-structure ←: in other words, %COARG is a coargument of the antecedent (c in (44)). According to the second constraint, %COARG is required to (possibly improperly) contain the reflexive (since there is a possibly empty path GF* through %COARG ending in ↑). According to the third constraint, the antecedent (the value of $\mathrm{GF}_{\mathrm{ant}}$, a in (44)) must be superior on the grammatical function hierarchy to %COARG. For conciseness and to allow reuse of this set of con-

straints by other lexical forms, we define the template SUPERIOR-ANT as encoding exactly this set of constraints:

(47) Definition of the template SUPERIOR-ANT:

$$\begin{array}{lll} \textsc{superior-ant} & \equiv & \%\textsc{coarg} = (\leftarrow \textsc{gf}) \\ & & (\%\textsc{coarg}\ \textsc{gf}^{*}) = \uparrow \\ & & \textsc{superior}(\rightarrow, \%\textsc{coarg}) \end{array}$$

This allows us to succinctly state the binding conditions on *himself* by means of the expression in (48), with a template call @SUPERIOR-ANT to the template defined in (47):

(48) Superior f-structures in the Minimal Complete Nucleus:

$$\begin{array}{l} ((\textsc{mcnpath} \uparrow) \quad \textsc{gf}_{\text{ant}} \quad) \\ \qquad\qquad @\textsc{superior-ant} \end{array}$$

The expression in (48) ranges over f-structures that constitute appropriate antecedents for *himself* in that they bear a superior grammatical function (as defined by the template @SUPERIOR-ANT) within the Minimal Complete Nucleus containing *himself* (as specified in the definition of MCNPATH).

It is easily seen from the topology of (40) that both *Alan* and *Bertie* are permissible antecedents. However, the expression in (41) picks out a single antecedent, and cannot be resolved to both at the same time. We therefore correctly predict that split antecedents are not possible with reflexives, as shown in (49).

(49) *Alan$_i$ told Bertie$_j$ about themselves$_{i+j}$

Now that we can refer to the f-structures that are syntactically suitable antecedents for a reflexive, it is possible to state the appropriate constraint on the $\mathscr{R}$ relation between the index of the reflexive and the index of its antecedent in terms of the expression in (48). This can be done by augmenting (48) with specification of the $\mathscr{R}$ relation. In (50), $\mathscr{R}$ relates the index of *himself* (which appears as the value of INDEX in its semantic structure $\uparrow_\sigma$) to the index of a superior f-structure within the Minimal Complete Nucleus. This constraint is specified in the lexical entry for *himself*.

(50) Positive binding constraint for *himself*:

$$\mathscr{R}((\uparrow_\sigma \textsc{index})) = (((\textsc{mcnpath} \uparrow) \quad \textsc{gf}_{\text{ant}} \quad)_\sigma \textsc{index})$$
$$@\textsc{superior-ant}$$

Returning to the sentence in (39), the f-structure for *Alan* appears within the Minimal Complete Nucleus relative to the f-structure of the reflexive pronoun, and stands in the appropriate superiority relation to the f-structure of the reflexive pronoun. Thus, if *Alan* serves as antecedent to the pronoun, as shown in (40), the conditions in (43) and (50) are met, and the binding relation is permitted.

8 PERSONAL PRONOUNS AND NEGATIVE CONSTRAINTS

We now turn to the more complex case of pronouns that are subject to negative constraints, such as English personal pronouns. As discussed in relation to example (5) above, when it comes to negative constraints we must deal with the fact that the constraint is not merely about antecedency, but about non-coreference. Antecedency implies coreference, but coreference can obtain even between elements that are not in an antecedency relation. If, therefore, we state a positive constraint defining a relation of antecedency, by implication we define a relation of coreference. But if we state a negative constraint ruling out a relation of antecedency, we do not necessarily rule out coreference.

We assume that all anaphoric indices are ultimately related to one non-anaphoric index, although the relation may not be direct. For example, consider the following sentence:

(51) Bertie said that he$_1$ thought that he$_2$ would win.

In (51), it may be that *Bertie*, *he*$_1$ and *he*$_2$ are all coreferent, and that *he*$_2$ takes *he*$_1$ as antecedent, and *he*$_1$ takes *Bertie* as antecedent. In this case, *he*$_2$ and *Bertie*, although coreferent, are not directly connected with one another via the $\mathscr{R}$ relation. Rather, they are related indirectly: $\mathscr{R}$ applied to the index of *he*$_2$ finds the index of *he*$_1$, and $\mathscr{R}$ applied to the index of *he*$_1$ finds the index of *Bertie*. While antecedency corresponds to a direct relation between the index of a pronoun and the index of its antecedent, coreference corresponds to the equivalence relation we get by taking the transitive, symmetric, reflexive closure of $\mathscr{R}$.[13] The class of discourse referents corresponding to this equivalence class of

[13] Observe that this applies to *intended* coreference only. Accidental coreference, as discussed in the binding literature, is presumably not reflected in $\mathscr{R}$.

indices is what Kamp and Reyle (1993, 235–236) refer to as $[x]_{\mathrm{K}}$ in their discussion of negative constraints: the class of discourse referents identified via equality with a given discourse referent x relative to a DRS K.

We choose as a representative of the equivalence class induced by (the closure of) $\mathscr{R}$ the first, non-anaphoric index. We therefore provisionally define the function $\mathscr{R}^*$ as in (52); a refinement to this definition will be necessary in the analysis of plural anaphora. This definition allows us to state negative binding constraints in terms of noncoreference, as required.

(52) Definition of $\mathscr{R}^*$, version 1 (to be amended):

$$\mathscr{R}^*(x) = \begin{cases} x & \text{if } \mathscr{R}(x) \text{ is undefined} \\ \mathscr{R}^*(\mathscr{R}(x)) & \text{otherwise} \end{cases}$$

$\mathscr{R}^*$ effectively follows the $\mathscr{R}$ path back from index to antecedent index, stopping only when it finds a non-anaphoric index: that is, an index without an antecedent. Note that $\mathscr{R}(x)$ is undefined for a non-anaphoric index, and the definition of $\mathscr{R}^*$ means that $\mathscr{R}^*(x)$ is x itself, if x is non-anaphoric. Choosing the first, non-anaphoric index as a representative of the coreference class is to some extent an arbitrary choice; presumably speakers do not always go back to the first mention of a new referent in a discourse. It would be possible to use instead the earliest occurrence within n sentences, but we assume this is a processing issue that we can legitimately abstract away from.

As discussed above, the English personal pronouns are subject to a negative constraint that refers to the Coargument Domain of the pronoun: the pronoun may not corefer with a superior coargument. As is standard, we define the Coargument Domain in terms of the path CoargPath:

(53) Coargument binding domain:

$$\textsc{CoargPath} \equiv \begin{array}{cc} \textsc{gf}^* & \textsc{gf} \\ \neg(\rightarrow \textsc{pred}) & \end{array}$$

Given this definition and the definition of the template @superior-ant in (47), the set of f-structures which may not bind a pronoun is the set of superior f-structures in the Coargument Domain, which can be referred to in the following way:

(54) Superior f-structures in the Coargument Domain:

$$((\textsc{CoargPath} \uparrow) \quad \textsc{gf}_{\textsc{ant}} \quad) \\ @\textsc{superior-ant}$$

The expression (CoargPath ↑) refers to the f-structures in the minimal domain containing ↑, the f-structure of the pronoun, which do not properly contain an f-structure with a PRED feature. This gives us the Coargument Domain. We can then refer to superior coarguments bearing the grammatical function $\textsc{gf}_{\textsc{ant}}$ within this domain by imposing the same off-path constraints as in (48). These are the f-structures with which the pronoun may not corefer.

Given the definition in (52), we can now formalise the negative constraint on English personal pronouns like *him* and *her* in the following way:

(55) Negative binding constraint for English personal pronouns:

$$\mathscr{R}^*((\uparrow_\sigma \textsc{index})) \neq \\ \mathscr{R}^*(((\textsc{CoargPath} \uparrow) \quad \textsc{gf}_{\textsc{ant}} \quad)_\sigma \textsc{index})) \\ @\textsc{superior-ant}$$

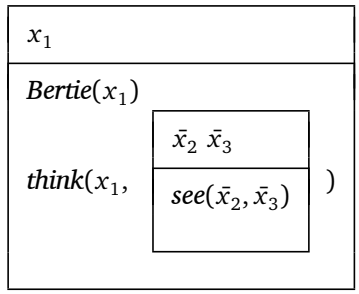
This constraint is specified in the lexical entry of personal pronouns such as *him*. It incorporates the expression in (54), which refers to superior coarguments of the personal pronoun, and it requires that the non-anaphoric index which is the antecedent (of the antecedent (of the antecedent…)) of the pronoun not be identical with any[14] non-anaphoric index introduced by, or serving as antecedent (of an antecedent (of an antecedent…)) to a superior coargument of ↑. The constraint ensures that non-coreference is enforced even when the immediate antecedents of two coargument pronouns are different, by following the $\mathscr{R}$ paths back to a non-anaphoric index, and ensuring that the two paths do not lead to the same index. The use of templates such as @SUPERIOR-ANT and path definitions such as MCNPATH and COARGPATH allows us to capture commonalities in binding requirements across all anaphoric elements within and across languages.

To illustrate the effect of the constraint in (55), consider example (5), repeated in (56) with its DRS, representing the monotonic meaning of the sentence.[15]

[14] Notice that the negation scopes over the disjunction over grammatical functions in the Coargument Domain, giving universal force.

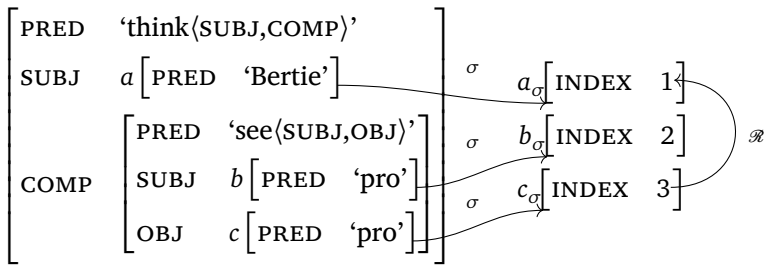
[15] We follow Maier (2009) in analyzing propositional attitudes as relations

(56) Bertie thought that he had seen him.

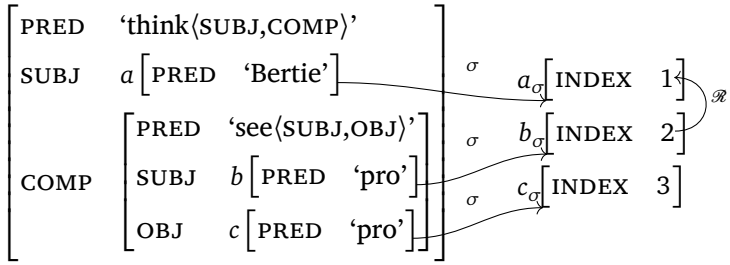


In (56), *he* may not serve as antecedent for *him*, since they are coarguments. *Bertie* may serve as antecedent for either *he* or *him*, but crucially may not serve as antecedent for both, since *he* and *him* may not be coreferent. The equation in (55) licenses the interpretations schematized in (57) and (58), both of which are possible, but rules out the interpretations schematized in (59) and (60), since these both involve coreference of coarguments.

(57) Bertie$_i$ thought that he$_j$ had seen him$_i$.

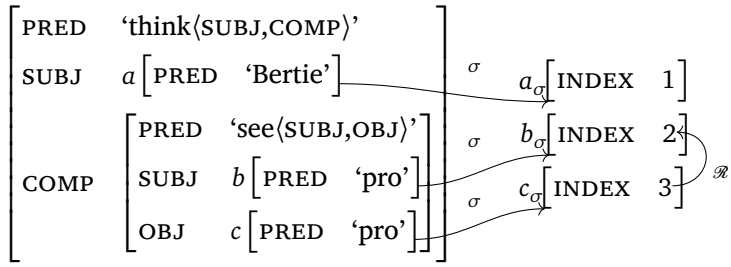


(58) Bertie$_i$ thought that he$_i$ had seen him$_j$.

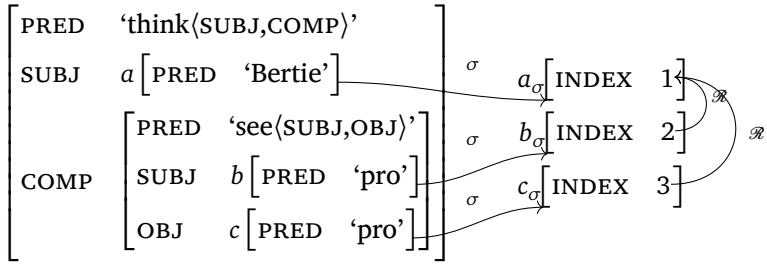


between individuals and DRSs. This would require an intensional language, but since attitudes are orthogonal to our concerns, we omit details here.

(59) *Bertie$_i$ thought that he$_j$ had seen him$_j$.



(60) *Bertie$_i$ thought that he$_i$ had seen him$_i$.



In this section, we have shown how to model the syntactic constraints imposed on the pragmatic interpretation of pronouns, using the projection architecture of LFG to constrain possible relations between indices and therefore the discourse referents that they introduce. In the next section, we show how plural pronouns complicate this picture, and we show that our model is able to handle these complications.

9 FORMALISING PLURAL ANAPHORS

9.1 *Complex indices and complex discourse referents*

As discussed above, the main complication that arises when we turn to plural pronouns is that we can no longer think purely in terms of relations between atomic indices and hence atomic discourse referents. The first fundamental change that we must make to the model presented above is to introduce a means of forming complex indices, associated with complex discourse referents, which can serve as antecedents for plural pronouns. Complex indices can be associated with

coordinated noun phrases: for example, the index for a coordinated phrase like *Alan and Bertie* is a complex index formed by combining the index for *Alan* with the index for *Bertie*. We can also form complex indices by combining the indices of contextually salient discourse referents; this is necessary in the analysis of split antecedency in examples like (61), where the antecedent of *them* is the complex index formed from the indices of the contextually salient discourse referents for *Alan* and *Bertie*.

(61) Alan_i told Bertie_j that Charlie_k admired them_{i+j}.

On the other hand, we do not need complex indices on lexical items, as simple indices are enough to capture e.g. binding by a group noun or a plural.[16] Thus, we assume that no lexical item introduces a complex discourse referent: these arise through phenomena such as coordination, split antecedence, etc.

Complex indices and discourse referents are formed by a mereological sum operator $\oplus$.[17] For discourse referents, this is what Kamp and Reyle (1993, Chapter 4) call *Summation* (see also Berman and Hestvik 1997, Section 3); for indices, this is similar to what Büring (2005, Section 9.3) calls an *index set*, a proposal with its roots in work by Lasnik (1989a). We do not go into formal details here, but intuitively this means that we no longer have only the atomic[18] indices $1, 2, \ldots$ and discourse referents $x_1, x_2, \ldots$, but also complex indices $1 \oplus 2, 3 \oplus 7, \ldots$ and discourse referents $x_1 \oplus x_2, x_3 \oplus x_7, \ldots$.

The notion of mereological sum is familiar from the literature on plurals (Link 1983). Adopting precisely that theory of plurals, we can easily make sure that complex discourse referents are properly interpreted. Recall from Section 5 that it is the ν function that lets us move from discourse referents to their inhabitants in a given state of the discourse. To make sure that we can do the same for complex

[16] We assume that an example like *The boys talked about him*, where *him* is one of the boys, exemplifies Pattern B as described in Section 4. As with the examples discussed there, we assume that such examples violate no syntactic binding constraints, though they may be unacceptable for nonsyntactic reasons.

[17] Technically, these are distinct domains with distinct sum operators, but we simplify matters here.

[18] An index i is atomic iff there are no two distinct indices such that their mereological sum equals i.

discourse referents, we introduce the axiom in (62), writing $\oplus^*$ for Link's sum operator on individuals.[19]

(62) Relation between complex discourse referents and their inhabitants:

$$\forall s \forall \delta \forall \delta' . \nu(s)(\delta \oplus \delta') = \nu(s)(\delta) \oplus^* \nu(s)(\delta')$$

That is, in all states, the inhabitant of the complex discourse referent $\delta \oplus \delta'$ is the sum of the inhabitants of the discourse referents δ and δ'. Note that the homomorphism from discourse referents to individuals is not (necessarily) an isomorphism, so that while non-atomic discourse referents map to non-atomic individuals, the converse is not necessarily true: as we already saw, a group noun will introduce an atomic discourse referent inhabited by a non-atomic individual.

9.2 *Complex indices and $\mathscr{R}^*$*

Above, we defined $\mathscr{R}^*$ as a recursive version of $\mathscr{R}$, as a way of moving from indices of anaphoric expressions back to indices with no antecedent, moving perhaps through one or more indices of anaphoric expressions on the way. But this assumed that all indices are atomic. Now that we have introduced complex indices, we must update our definition of $\mathscr{R}^*$ accordingly. Notice that we assume that all complex indices are ultimately constructed out of atomic indices. We can therefore define a function $\textsc{atoms}(i)$ which returns the set of atomic indices that make up i. With this in place, we revise the definition of $\mathscr{R}^*$ as in (63):

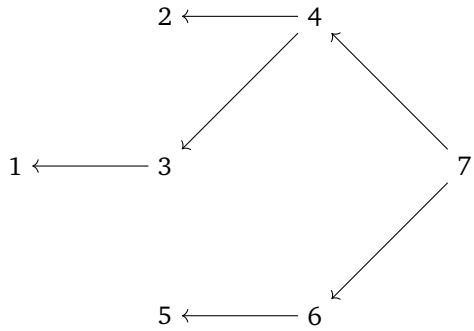
(63) Definition of $\mathscr{R}^*$ (final; includes additional condition for complex indices):

$$\mathscr{R}^*(x) = \begin{cases} x & \text{if } x \text{ is atomic and } \mathscr{R}(x) \text{ is undefined} \\ \mathscr{R}^*(\mathscr{R}(x)) & \text{if } x \text{ is atomic and } \mathscr{R}(x) \text{ is defined} \\ \oplus\{\mathscr{R}^*(y) \mid y \in \textsc{atoms}(x)\} & \text{otherwise (i.e. if } x \text{ is non-atomic)} \end{cases}$$

[19] We can achieve the same result in a more general setting by requiring a homomorphism from the algebra of discourse referents to the algebra of individuals.

In words, atomic indices are treated as before: we follow the antecedency path as far as possible. For non-atomic indices we simply apply the function to their atomic parts and take the sum of the results. To see how this works, consider the diagram in (64).

(64)



This diagram represents a situation in which there are seven indices, four of them belonging to anaphoric expressions, with the following $\mathscr{R}$ relations: $\mathscr{R}(7) = 4 \oplus 6$; $\mathscr{R}(4) = 2 \oplus 3$; $\mathscr{R}(6) = 5$; $\mathscr{R}(3) = 1$. This situation is exemplified by the following text:

(65) John$_1$ came in and sat down. Paul$_2$ sat down next to him$_3$, and they$_4$ got out their instruments. Next, George$_5$ arrived, and he$_6$ sat down at the piano. They$_7$ all started to sing.

x_1 x_2 $\bar{x}_3$ $\bar{x}_4$ x_5 $\bar{x}_6$ $\bar{x}_7$
John(x_1) *come-in*(x_1) *sit-down*(x_1) *Paul*(x_2) *sit-down-next-to*$(x_2, \bar{x}_3)$ *get-out-instruments*$(\bar{x}_4)$ *George*(x_5) *arrive*(x_5) *sit-down-at-piano*$(\bar{x}_6)$ *sing*$(\bar{x}_7)$

$, \mathscr{R} : 7 \mapsto 4 \oplus 6,\ 4 \mapsto 2 \oplus 3,\ 6 \mapsto 5,\ 3 \mapsto 1$

We get $\mathscr{R}^*(7)$ from the given $\mathscr{R}$ in the following way: By the second clause of (63), $\mathscr{R}^*(7) = \mathscr{R}^*(4 \oplus 6)$. By the third clause $\mathscr{R}^*(4 \oplus 6) = \mathscr{R}^*(4) \oplus \mathscr{R}^*(6)$. By the two first clauses, $\mathscr{R}^*(6) = 5$, whereas $\mathscr{R}^*(4) =$

$2 \oplus 3$, and $\mathscr{R}^*(3) = 1$. So we get $\mathscr{R}^*(7) = \mathscr{R}^*(4 \oplus 6) = 1 \oplus 2 \oplus 5$ (since mereological sum is associative).

9.3 *Reflexives with plural antecedents*

In the case of the positive constraint on English reflexive pronouns, plurality has little effect on the generalisations. A reflexive pronoun must be coreferent with an antecedent in its Minimal Complete Nucleus, whether it is singular or plural. Partial coreference is not possible: for example, one cannot say the following, to mean that Bertie likes himself and one or more others:

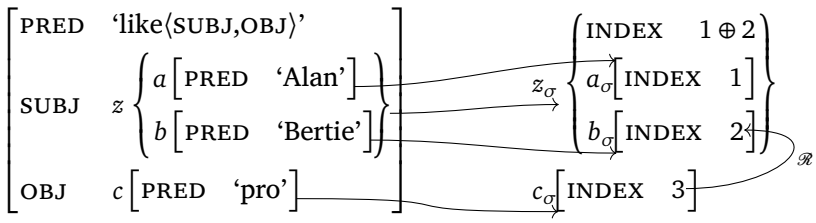
(66) *Bertie likes themselves.

Likewise, one cannot say the following, to mean that Alan and Bertie like one of either Alan or Bertie:

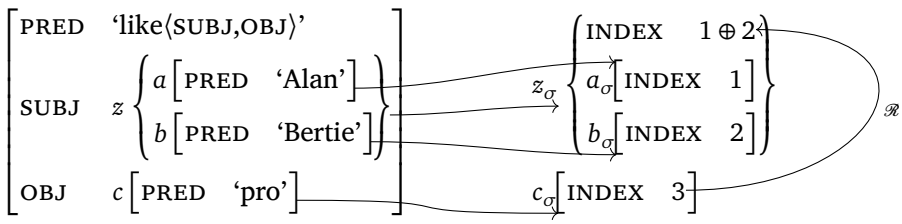
(67) *Alan and Bertie like himself.

A reflexive pronoun must take a single (possibly complex, but not split) index as its antecedent. Given our analysis of complex indices in terms of mereological sums, the plural case falls directly out of equation (50), as shown in (68)–(69). Note that the f-structure for a coordinate structure like *Alan and Bertie* is a set (labeled z), and the conjuncts are elements of the set (labeled a and b).

(68) *Alan$_i$ and Bertie$_j$ like himself$_j$.



(69) Alan$_i$ and Bertie$_j$ like themselves$_{i+j}$.



The f-structure z is the only f-commanding GF in the Minimal Complete Nucleus relative to c. The only licit antecedent is therefore (z_σ INDEX), which is $1 \oplus 2$. The f-structures for *Alan* and *Bertie* are not syntactically suitable antecedents – in particular, they do not f-command the pronoun – and so their indices 1 and 2 are not individually available as antecedents.

9.4 *Pronouns with split antecedents*

With (63) in place, it is unproblematic to account for examples like (61), repeated here, making use of the negative constraint in (55).

(70) Alan_1 told Bertie_2 that Charlie_3 admired them_4.

Let us check that the resolution $\mathscr{R}^*(4) = 1 \oplus 2$ is valid. If we assume that these are the first occurrences of *Alan*, *Bertie*, and *Charlie* in the discourse, we get $\mathscr{R}^*(1) = 1, \mathscr{R}^*(2) = 2$, $\mathscr{R}^*(3) = 3$, and $\mathscr{R}^*(4) = 1 \oplus 2$. By (55), $\mathscr{R}^*(4)$ must be different from $\mathscr{R}^*$ applied to any index projected from a superior element in its binding domain, which is the Coargument Domain. The only superior coargument is *Charlie*, so $\mathscr{R}^*(4)$ must be different from $\mathscr{R}^*(3)$, which it is.

9.5 *Comparison with other approaches*

To our knowledge, Berman and Hestvik (1997) is the most recent attempt to deal with the binding patterns of plural pronouns. Besides offering a more precise formalization, Berman and Hestvik (1997) also discuss and improve upon certain aspects of Lasnik (1989a), Seeley (1993), and Fiengo and May (1994). Therefore, we only compare our approach to Berman and Hestvik (1997) here.

The main empirical difference between our approach and that of Berman and Hestvik (1997) concerns Principle B effects in ditransitives. They claim that it is ungrammatical for a pronoun to corefer with the *sum* of its superior coarguments (Pattern D above) and set their theory up accordingly. Our theory instead predicts that a pronoun must be non-coreferent with *each* of its superior coarguments, meaning that Pattern D is grammatical.

As we noted in our discussion of Pattern D in Section 4, the empirical evidence is unclear. Note that both approaches make the same predictions about standard, monotransitive cases like *John likes him*,

because the sum of superior coarguments of the pronoun in such cases just is the single superior coargument.

The theories also differ in the predictions they make about examples where a pronoun corefers with one of several superior coarguments. Such examples are in fact unacceptable, and are judged as such by Berman and Hestvik themselves.

(71) *John$_i$ told Mary$_j$ about her$_j$/him$_i$. (Berman and Hestvik 1997, 25)

This pattern is incorrectly classified as grammatical by the restriction on coreference proposed by Berman and Hestvik (1997, 22): "the restriction on CR.PRO [coreference resolution of pronouns] is simply that no DRS-equivalent of a potential resolving discourse referent for a pronoun may be identical to the set of discourse referents that c-command the pronoun within its binding domain". Berman and Hestvik appear not to have noticed that this runs counter to their judgements about examples like (71). Their theory could probably be amended by making the generalization (and the corresponding formalization) disjunctive ("or with a single c-commanding discourse referent"). Similarly, should further empirical investigation reveal that Pattern D is indeed ungrammatical, our theory could be amended with the extra constraint in (72).

(72) Additional negative condition for plural pronouns, requiring noncoreference with the sum of the coarguments:

$$\mathscr{R}^*((\uparrow_\sigma \textsc{index})) \neq \oplus\{x \mid x = \mathscr{R}^*((((\textsc{CoargPath} \uparrow) \; \textsc{gf}_{\textsc{ant}} \; @\textsc{superior-ant})_\sigma \textsc{index}))\}$$

However, at this stage, we do not see any way of ruling out Pattern D and the other illicit binding patterns for pronouns by means of a single, nondisjunctive constraint. That is, should Pattern D turn out to be ungrammatical, it seems that the negative binding constraints on (plural) pronouns would have to be essentially disjunctive.

10 CONCLUSION

The interaction of syntactic and pragmatic constraints on pronominal reference provides a challenge for any model of grammar. Our

approach offers an integrated account in which syntactic and pragmatic factors jointly constrain binding possibilities. In our model, binding theory is stated in terms of syntactic constraints on pragmatic anaphora resolution. The modular grammatical architecture of LFG provides a natural setting for this integration, with its clean separation of syntactic, semantic, and pragmatic components of the grammar. In this, our analysis represents a step forward from the most recent in-depth work on binding of plural anaphora, the work of Berman and Hestvik (1997), who present an approach involving rewriting of Government and Binding-style phrase structure trees into DRSs. We also provide for the first time a full formal treatment of coreference relations and negative binding constraints in an LFG setting. Our analysis crucially relies on the Partial Compositional Discourse Representation Theory of Haug (2014b), with its explicit separation between the semantic and pragmatic contributions of anaphoric elements.

Regarding the empirical data for plural anaphora, we have identified four possible patterns of inclusion between the index of a pronoun and its antecedent, some of which have been subject to varying grammaticality judgements in previous literature. Our formal analysis classifies these patterns as syntactically wellformed, and we anticipate that further research will uncover other factors, such as lexical and contextual factors, to explain unacceptable instances.

Further potential for our analysis includes its extension to modelling constraints on resumptive pronouns (Asudeh 2011, 2012) and null pronouns e.g. in anaphoric control constructions; PCDRT has already been extended to deal with the anaphoric relations inherent in partial control constructions (Haug 2014a; see also Haug 2013 and Belyaev and Haug 2014).

ACKNOWLEDGMENTS

We are grateful to the audience at LFG17 in Konstanz for comments on an early version of this paper, particularly to Ash Asudeh and Geoff Pullum. We are also grateful to Jamie Findlay, Adam Przepiórkowski, and three anonymous JLM reviewers for comments on a later version. This article was finished while the two first authors were on sabbatical at the Centre for Advanced Study at the Norwegian Academy of Science and Letters. We gratefully acknowledge their support.

REFERENCES

Ash ASUDEH (2011), Towards a unified theory of resumption, in Alain ROUVERET, editor, *Resumptive pronouns at the interfaces*, pp. 121–187, Benjamins, Amsterdam.

Ash ASUDEH (2012), *The logic of pronominal resumption*, Oxford University Press, Oxford.

Ash ASUDEH and Gianluca GIORGOLO (2012), Flexible composition for optional and derived arguments, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG12 Conference*, pp. 64–84, CSLI Publications, Stanford, CA.

Ash ASUDEH, Gianluca GIORGOLO, and Ida TOIVONEN (2014), Meaning and valency, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG14 Conference*, pp. 68–88, CSLI Publications, Stanford, CA.

David I. BEAVER (1992), The kinematics of presupposition, in Paul DEKKER and Martin STOCKHOF, editors, *Proceedings of the Eighth Amsterdam Colloquium*, ILLC, University of Amsterdam.

David I. BEAVER (2002), Presupposition projection in DRT: a critical assessment, in David I. BEAVER, editor, *The construction of meaning*, pp. 23–43, CSLI Publications, Standford, CA.

Oleg I. BELYAEV and Dag T. T. HAUG (2014), Pronominal coreference in Ossetic correlatives and the syntax-semantics interface, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG14 Conference*, pp. 89–109, CSLI Publications, Stanford, CA.

Stephen BERMAN and Arild HESTVIK (1997), Split antecedents, noncoreference, and DRT, in Hans BENNIS, Pierre PICA, and Johan ROORYCK, editors, *Atomism and Binding*, pp. 1–29, Foris, Dordrecht.

Tina BÖGEL, Miriam BUTT, Ronald M. KAPLAN, Tracy Holloway KING, and John T. MAXWELL III (2009), Prosodic phonology in LFG: a new proposal, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG09 Conference*, pp. 146–166, CSLI Publications, Stanford, CA.

Joan BRESNAN, Ash ASUDEH, Ida TOIVONEN, and Stephen WECHSLER (2016), *Lexical-functional syntax*, Wiley-Blackwell, Oxford, second edition. First edition by Joan Bresnan, 2001, Blackwell.

Joan BRESNAN, Per-Kristian HALVORSEN, and Joan MALING (1985), Logophoricity and bound anaphors, unpublished manuscript, Department of Linguistics, Stanford University.

Daniel BÜRING (2005), *Binding theory*, Cambridge University Press, Cambridge.

Gennaro CHIERCHIA (2004), Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface, in Adriana BELLETTI, editor, *Structures and*

beyond: the cartography of syntax structures, vol. 3, pp. 39–103, Oxford University Press, Oxford.

Noam CHOMSKY (1973), Conditions on transformations, in Stephen ANDERSON and Paul KIPARSKY, editors, *A Festschrift for Morris Halle*, Holt, Rinehart & Winston, New York.

Noam CHOMSKY (1981), *Lectures on Government and Binding: the Pisa lectures*, Foris, Dordrecht.

Mary DALRYMPLE (1993), *The syntax of anaphoric binding*, CSLI Publications, Stanford, CA.

Mary DALRYMPLE (2001), *Lexical Functional Grammar*, Academic Press, San Diego, CA.

Mary DALRYMPLE and Louise MYCOCK (2011), The prosody-semantics interface, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG11 Conference*, pp. 173–193, CSLI Publications, Stanford, CA.

Mary DALRYMPLE and Irina NIKOLAEVA (2006), Syntax of natural and accidental coordination: evidence from agreement, *Language*, 82(4):824–849.

Mary DALRYMPLE and Irina NIKOLAEVA (2011), *Objects and information structure*, Cambridge University Press, Cambridge.

Yehuda N. FALK (2001), *Lexical-Functional Grammar: an introduction to parallel constraint-based syntax*, CSLI Publications, Stanford, CA.

Robert FIENGO and Robert MAY (1994), *Indices and identity*, Linguistic Inquiry Monographs, The MIT Press, Cambridge, MA.

Gianluca GIORGOLO and Ash ASUDEH (2011), Multimodal communication in LFG: gestures and the correspondence architecture, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG11 Conference*, pp. 257–277, CSLI Publications, Stanford, CA.

Dag T. T. HAUG (2013), Partial control and the semantics of anaphoric control in LFG, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG13 Conference*, pp. 274–294, CSLI Publications, Stanford, CA.

Dag T. T. HAUG (2014a), The anaphoric semantics of partial control, *Proceedings of SALT*, 24:213–233.

Dag T. T. HAUG (2014b), Partial dynamic semantics for anaphora: compositionality without syntactic coindexation, *Journal of Semantics*, 31(4):457–511, first published online August 24, 2013. DOI: 10.1093/jos/fft008.

Irene R. HEIM (1982), *The semantics of definite and indefinite noun phrases*, Ph.D. thesis, University of Massachusetts, Amherst.

James HIGGINBOTHAM (1983), Logical form, binding, and nominals, *Linguistic Inquiry*, 14(3):395–420.

Ray JACKENDOFF (1992), Mme. Tussaud meets the binding theory, *Natural Language and Linguistic Theory*, 10(1):1–32.

Hans KAMP and Uwe REYLE (1993), *From discourse to logic*, Kluwer Academic Publishers, Dordrecht.

Hans KAMP, Josef VAN GENABITH, and Uwe REYLE (2011), Discourse Representation Theory, in Dov M. GABBAY and Franz GÜNTHNER, editors, *Handbook of philosophical logic*, pp. 125–394, Springer, Dordrecht, second edition.

Ronald M. KAPLAN (1989), The formal architecture of Lexical-Functional Grammar, in Chu-Ren HUANG and Keh-Jiann CHEN, editors, *ROCLING II: Proceedings of the Computational Linguistics Conference*, pp. 3–18, The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Tapei, also published in *Journal of Information Science and Engineering 5* (1989), pp. 305–322, and in *Formal issues in Lexical-Functional Grammar*, ed. Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III and Annie Zaenen, CSLI Publications, 1995, pp. 7–27.

Ronald M. KAPLAN and Joan BRESNAN (1982), Lexical-Functional Grammar: a formal system for grammatical representation, in Joan BRESNAN, editor, *The mental representation of grammatical relations*, pp. 173–281, MIT Press, Cambridge, MA.

Edward L. KEENAN and Bernard COMRIE (1977), Noun phrase accessibility and universal grammar, *Linguistic Inquiry*, 8(1):63–99.

Paul KIPARSKY (2002), Disjoint reference and the typology of pronouns, in Ingrid KAUFMANN and Barbara STIEBELS, editors, *More than words: a Festschrift for Dieter Wunderlich*, pp. 179–226, Akademie Verlag, Berlin.

Miltiadis KOKKONIDIS (2008), First-order Glue, *Journal of Logic, Language and Information*, 17:43–68.

Howard LASNIK (1989a), On the necessity of binding conditions, in *Essays on anaphora*, Kluwer, Dordrecht.

Howard LASNIK (1989b), On two recent treatments of disjoint reference, in *Essays on anaphora*, Kluwer, Dordrecht.

Howard LASNIK (1989c), Remarks on coreference, in *Essays on anaphora*, Kluwer, Dordrecht.

Howard LASNIK (1989d), A selective history of modern binding theory, in *Essays on anaphora*, Kluwer, Dordrecht.

Godehard LINK (1983), The logical analysis of plurals and mass terms: a lattice-theoretical approach, in Rainer BÄUERLE, Chirstoph SCHWARZE, and Arnim VON STECHOW, editors, *Meaning, use and the interpretation of language*, pp. 303–323, de Gruyter, Berlin.

Emar MAIER (2009), Presupposing acquaintance: a unified semantics for *de dicto*, *de re* and *de se* belief reports, *Linguistics and Philosophy*, 32:429–474.

Reinhard MUSKENS (1996), Combining Montague Semantics and Discourse Representation, *Linguistics and Philosophy*, 19:143–186.

Louise MYCOCK and John J. LOWE (2013), The prosodic marking of discourse functions, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG13 Conference*, pp. 440–460, CSLI Publications, Stanford, CA.

Rick NOUWEN (2003), *Plural pronominal anaphora in context*, Ph.D. thesis, Utrecht Institute of Linguistics OTS.

Karen PARK (2012), *The selective properties of verbs in reflexive constructions*, D.Phil. thesis, University of Oxford.

Tanya REINHART and Eric REULAND (1993), Reflexivity, *Linguistic Inquiry*, 28:178–187.

T. Daniel SEELEY (1993), Binding plural pronominals, in Katherine BEALS, Gina COOKE, David KATHMAN, Sotaro KITA, Karl-Erik MCCULLOUGH, and David TESTEN, editors, *CLS29: papers from the 29th regional meeting of the Chicago Linguistic Society, vol. II*, pp. 305–317, University of Chicago, Chicago.

Thomas WASOW (1972), *Anaphoric relations in English*, Ph.D. thesis, Massachusetts Institute of Technology.

Sets, heads, and spreading in LFG

Avery D. Andrews
Australian National University

ABSTRACT

Lexical Functional Grammar (LFG) uses abstract syntactic representations (f-structures) that tend to provide less hierarchical structure for certain constructions than those employed in other formal frameworks. This produces some good results, such as a very straightforward account of feature-sharing between phrases and their heads, but also certain difficulties, especially in cases where the semantic interpretation seems to be determined by the hierarchical c-structure rather than the flatter f-structure. These are unproblematic for all other major generative frameworks, but have been troublesome for standard versions of LFG.

Here I will consider two such cases: scoping adjectival modification in noun phrases; and Romance 'complex' (or 'restructuring') predicates. Problems with the semantic interpretation of these constructions were first discussed by Andrews (1983) and Alsina (1997), respectively, and by others subsequently. Both constructions exhibit the problem of apparent concentricity, and a fully satisfactory and accepted LFG solution has not yet been found. My proposal is to use the hybrid objects and distribution convention of Dalrymple and Kaplan (2000), but with singleton rather than multi-member sets, along with a facility to stipulatively suppress distribution in individual constructions. This provides an analysis which explains scope-determination and helps with certain other problems, with far less change to the theory than in previous attempts such as Andrews and Manning (1993, 1999).

Keywords: hybrid objects, complex predicates, glue semantics, attribute spreading

1 INTRODUCTION

LFG has traditionally proposed relatively flat covert structures (f-structures) for a variety of constructions, such as adjectival modification and 'restructuring' complex predicates, which in most other frameworks are analysed as having hierarchical covert structures, usually binary branching ones. This leads to some problems for LFG that do not arise in other frameworks: most importantly, LFG does not provide an explanation for the apparent effects of concentric constituent structure on semantic interpretation; and LFG has problems implementing the associated morphological marking.

Andrews and Manning (1993, 1999) proposed to address these problems by means of substantial modifications to the LFG architecture, but those approaches, slightly different from each other, proved difficult to generalize to other phenomena, and did not recruit many followers. Here I will propose another and considerably simpler solution, based largely on machinery that LFG already uses, or that has at least some provisional acceptance for independent reasons. The core notions are those of hybrid object and distributive versus nondistributive attributes from Dalrymple and Kaplan (2000); another is to use the filtering properties of glue in place of traditional Completeness and Coherence. This was suggested as a possibility in some of the papers in Dalrymple (1999), and later by Kuhn (2001), and is accepted by Asudeh *et al.* (2014) and Lowe (2015). This proposal also requires minor additions to the formalism, along with changes to some familiar analyses (such as that of attributive adjectives) and to the default annotation rules.

In the next section, I will develop the basic theoretical ideas we will need; and in the third section I will present the treatment of modal and intersective adjectives, capturing the essential points from Andrews (1983) and Andrews and Manning (1993).[1] I will also analyse in LFG some material on agreement discrepancies that has recently been analysed in the Minimalist Program by Pesetsky (2013), Landau (2016) and Puškar (2017). In the fourth section, I will consider restructuring predicates in Catalan, where there is both a problem of scope in-

[1] We omit a treatment of what appear to be asyndetically coordinated adjectives, as in *a ruthless, unscrupulous property developer*, because analysing these requires a glue analysis of coordinate structures, taking us too far afield.

terpretation and one of form-determination. Although Catalan seems to be generally representative of the southern Romance languages, Alsina (1996, 1997) and Solà (2002) provide evidence that shows that the traditional LFG analysis of these constructions in Romance languages is not fully satisfactory. I conclude this section with a brief discussion of Hindi/Urdu causatives, as discussed by Lowe (2015), which are similar to Romance restructuring, but with the ordering reversed. Lowe analyses many important aspects of these constructions successfully within fully standard LFG + glue, and furthermore accomplishes the onerous task of carefully and cogently critiquing all previous analyses of restructuring complex predicates, but does not take on either scoping or form-determination.

2 HYBRID OBJECTS, DISTRIBUTION AND UNDERSHARING

Here we introduce the relatively new formal ideas we will need, hybrid objects and distribution, and the more recent proposal that I will call 'undersharing'. But glue semantics as presented in Dalrymple (2001) (the 'new glue' version) will be assumed, and not described here.

2.1 *Distribution vs. 'sharing'*

The notion of distributive attribute was introduced by Bresnan *et al.* (1985), and was further developed by Kaplan and Maxwell (1988). Distributive attributes, when attributed to a set in an f-structure, are in effect attributed to all the members of that set, and vice versa, allowing for the satisfaction of the Completeness and Coherence Constraints in examples such as *John bought and read the book*.

The formulation of distribution that we shall assume is from Dalrymple (2001, p. 158), and is slightly different from earlier ones such as Dalrymple and Kaplan (2000):

(1) For any distributive attribute A and set s, $A(s) = V$ iff $\forall f \in s$, $A(f) = V$.

To see how this works, consider a structure such as (2) below, where the attribute F is distributive, and the outer square brackets signify that the entire structure is actually a 'hybrid object' as we discuss in the next subsection, with both set-members, and, possibly, attributes:

(2) $\left[\left\{\left[\text{F}\quad Y\right]\atop\left[\text{F}\quad Z\right]\right\}\right]$

As long as nothing ascribes any F-value to the entire structure, it is possible that $Y \neq Z$. But if something ascribes a value X for F to the entire structure, then (2) must become more highly specified as indicated in (3) below (X is to be read as a shared value, rather than multiple copies, and the issue of whether X should or should not be written at the top level will be discussed shortly):

(3) $\left[\begin{array}{l}\text{F}\quad X\\ \left\{\left[\text{F}\quad X\right]\atop\left[\text{F}\quad X\right]\right\}\end{array}\right]$

And if this is impossible, due to Y and Z being contradictory, then there is no solution: there is no well-formed sentence structure that includes the f-structure. This is exactly the effect we want in coordinate structures, where grammatical relations are sometimes shared and sometimes not:

(4) a. Mary praised Bill and criticized John

b.
$$\left[\begin{array}{ll}\text{SUBJ} & \left[\text{PRED}\quad \text{‘Mary’}\right]\\ \text{CONJ} & \text{AND}\\ \left\{\begin{array}{l}\left[\begin{array}{ll}\text{SUBJ} & [\ ]\\ \text{PRED} & \text{‘praise(SUBJ, OBJ)’}\\ \text{TENSE} & \text{PAST}\\ \text{OBJ} & \left[\text{PRED}\quad \text{‘Bill’}\right]\end{array}\right]\\ \left[\begin{array}{ll}\text{SUBJ} & [\ ]\\ \text{PRED} & \text{‘criticize(SUBJ, OBJ)’}\\ \text{TENSE} & \text{PAST}\\ \text{OBJ} & \left[\text{PRED}\quad \text{‘John’}\right]\end{array}\right]\end{array}\right\}\end{array}\right]$$

In this case, SUBJ is supposed to be shared and OBJ is not, but other possibilities are both or neither:

(5) a. Mary praised Bill and Susan praised John.

b. Mary (both) praised and criticized John.

The distributivity convention (1) handles this and also other issues associated with coordinate structures; whereas the CONJ attribute in (4b) is nondistributive, and so is not shared amongst the conjuncts.

A further property of (1) is that if the values of F in the set members are specified as being the same by virtue of their internal structure, then this becomes the value of F for the entire structure as well, for the satisfaction of constraining equations. In the case of grammatical relations in coordinate structures, this will never happen due to the Predicate Indexing convention (all instances of PRED-values are taken as distinct, even if they represent the same choice from the lexicon), but it can occur for ordinary feature-values. In effect, distribution works the same way for defining specifications (those that impose a feature-value) applied to the whole and for constraining specifications (those that check that something else has put a given value somewhere).

Formulating the Coherence Constraint for the representation of (4b) is problematic. In the structures such as example (30) in Dalrymple (2001, p. 373), the distributed GFs are not explicitly represented at the upper level, perhaps on the basis that they are not 'really' present there, but are only 'virtually' present by the formulation of the definition (1), so that Coherence will work as usual. This can work for coordinate structures, since the lexical items calling for the grammatical functions are always located in the set members rather than in the whole structure. But in our analysis of complex predicates, grammatical relation attributes will be scattered across the levels of the set-inclusion structure, so we need to say something definite about this situation. One possibility would be to elaborate the definition of Coherence to deal with this; a simpler way is to dispense with the Coherence and Completeness Constraints in their original form, and let glue assembly do their work, as has been occasionally suggested since Kuhn (2001) if not before, and is accepted by Lowe (2015, p. 426).[2]

However, whether or not we abandon Completeness and Coherence, we have another problem with coordinate structures: the 'resource deficit' discussed by Dalrymple (2001, pp. 377–378) and Asudeh and Crouch (2002). The meaning resource provided by the subject in (4) needs to be consumed by two verbs, whereas by lin-

[2]The representational issue is addressed in greater detail in Appendix A.

ear logic, when one verb uses it, it is gone, and not available to the other. Asudeh and Crouch propose a solution that is notationally very complex, but works, and can be provisionally accepted here.

The behaviour of distribution when the set is a singleton has been somewhat overlooked. A distributive feature will always be distributed, and therefore in effect shared. In (6a), for example, X is the value of F in every member of the hybrid object's set, so (6a) comes out identical in its properties to (6b):

(6) a. $\left[\left\{\left[\mathrm{F} \quad X\right]\right\}\right]$

b. $\left[\begin{array}{l}\mathrm{F} \quad X \\ \left\{\left[\mathrm{F} \quad X\right]\right\}\end{array}\right]$

Distribution therefore produces effects very similar to the sharing of attributes used by Andrews and Manning (1993, 1999), but in a more limited way, and without any fundamental change to the formal framework beyond what is independently proposed for coordinate structures.

2.2 *Hybrid objects and 'undersharing'*

Hybrid objects were originally proposed by John Maxwell and introduced into the LFG literature by Dalrymple and Kaplan (2000, see esp. p. 778). A hybrid object is an f-structure that has not only members and distributive attributes, but can also have 'nondistributive' attributes that apply to the entire structure but that do not obey the distribution convention (1).

Person and number were the most important originally motivated nondistributive attributes. These are motivated by coordinate structures such as *José y yo* in Spanish, where both conjuncts are singular, but the whole NP is plural; and where one conjunct is first person, the other third, while the whole is first person:

(7) José y yo hablamos.
José and I talk.1PL(PRES or PRET)
'Jose and I talk/talked.'

Dalrymple and Kaplan propose the following f-structure for the NP (they omit the CONJ feature without discussion):

(8)
$$\left[\begin{array}{l}\left\{\left[\begin{array}{ll}\text{PRED} & \text{‘José’}\\ \text{PERS} & 3\\ \text{NUM} & \text{SG}\end{array}\right]\right.\\ \left.\left[\begin{array}{ll}\text{PRED} & \text{‘pro’}\\ \text{PERS} & 1\\ \text{NUM} & \text{SG}\end{array}\right]\right\}\\ \begin{array}{ll}\text{PERS} & 1\\ \text{NUM} & \text{PL}\end{array}\end{array}\right]$$

The values of NUM and PERS for the entire structure do not appear in all of the individual conjuncts, although the PERS-value does appear in one of them.

For Dalrymple and Kaplan's purposes, it is at least plausible that there is a universal classification of features into distributive and nondistributive (although, as we shall see, this is not entirely free of problems), but for the wider application of distribution that we are attempting here, this is unfortunately not possible. Rather, it seems necessary to stipulate on a construction-specific basis that certain features are not distributed.

Although it is not the only possibility, I propose that:

(9) a. Certain attributes, particularly ADJUNCT and CONJ (and possibly PRED) are universally non-distributive. In situations where they might appear to be distributive, some other analysis is correct, such as the use of functional uncertainty (no such cases are suggested here).

b. Other attributes are distributive by default, but these can be blocked from distribution by what I will call an 'undersharing' specification, as detailed below. In such cases, there is plentiful and overt positive evidence that the undershared attribute is behaving differently from the ones that are behaving distributively.

'Undersharing' as notated and used here is an innovation of this paper; but construction-specific stipulation of distributivity for attributes was suggested by Belayev *et al.* (2015).

3 ATTRIBUTIVE ADJECTIVES AND NP STRUCTURE

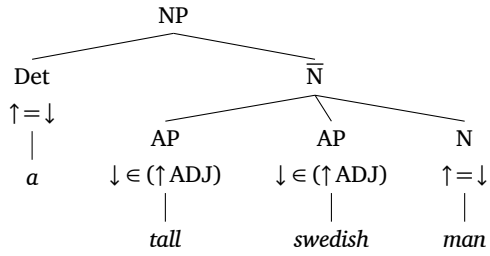
We now consider the relative scope of adjectival modifiers, first discussed by Andrews (1983) as an objection to the flat structure analyses

of Jackendoff (1977). This material, entirely unproblematic in most generative frameworks, was treated in the heavily modified version of LFG used in Andrews and Manning (1993), but not in the somewhat differently modified version of Andrews and Manning (1999). In the first subsection we consider the interactions of relative order and scope in English, with special attention to 'modal' adjectives such as *former* and *alleged*; in the second, we provide an analysis; in the third we discuss coordination and the need for undersharing stipulations; and in the fourth, we discuss the phenomenon of 'agreement mismatches' in certain other languages that provides additional motivation for the present approach.

3.1 *Adjectives and scope*

LFG has generally followed the 'flat structure' approach to adjectival modifiers advocated by Jackendoff (1977), e.g. Dalrymple (2001, pp. 256–257). The adjectives are introduced in APs whose f-structure correspondents are members of the set-valued attribute ADJUNCTS, yielding an annotated c-structure as follows for *a tall Swedish man*:[3,4]

(10) a.



b.

$$\begin{bmatrix} \text{SPEC} & \text{INDEF} \\ \text{ADJUNCTS} & \left\{ \begin{bmatrix} \text{PRED} & \text{'tall'} \end{bmatrix}, \begin{bmatrix} \text{PRED} & \text{'Swedish'} \end{bmatrix} \right\} \\ \text{PRED} & \text{'man'} \end{bmatrix}$$

This flat f-structure works well for intersective adjectives, as treated in considerable detail by Dalrymple. It can be extended to at least some

[3] Dalrymple (2001, p. 257) omits from the structure the topmost NP layer with the determiner.

[4] Note that the set-values of ADJUNCTS have never been argued to be hybrid objects, so we seem to have an implicit distinction between hybrid objects and 'pure sets', which would not be able to have nondistributive attributes.

subsectives, such as *skillful*, by treating them as taking an unexpressed *as*-argument. This is usually supplied by the head noun when the adjective is in attributive position, but is fundamentally always supplied by context, most obviously so when the adjective is predicative:[5]

(11) a. Brett is a skillful surgeon, but not much of a pilot.
b. Wow, he's skillful! [meaning: as a surgeon, watching Brett in the operating theater and implying nothing about his piloting skills]
c. Can we find any good linguists? [meaning: good at basketball, for an interdepartmental tournament][6]

This analysis fails to give a fully satisfactory account of 'modal' adjectives such as *former* and *alleged*, because, although Dalrymple's glue treatment works when there are no other modifiers, such as *former* in *former senator*, it doesn't account for the effect of ordering on interpretation when there are multiple modifiers:

(12) a. He is an unscrupulous former property-developer.
b. He is a former unscrupulous property-developer.

The first characterizes his career as a developer as having existed in the past, but his unscrupulousness as persisting, while the second locates both in the past, so that he could well now be a comprehensively reformed character. We also note that *He is a formerly unscrupulous property developer* means that he's still a developer, but is no longer an unscrupulous one. When *former* is replaced by its adverbial variant, the attribution to past time applies only to the adjective, not the entire adj + noun combination (as is captured by Dalrymple's analysis of adverbs modifying attributive adjectives).

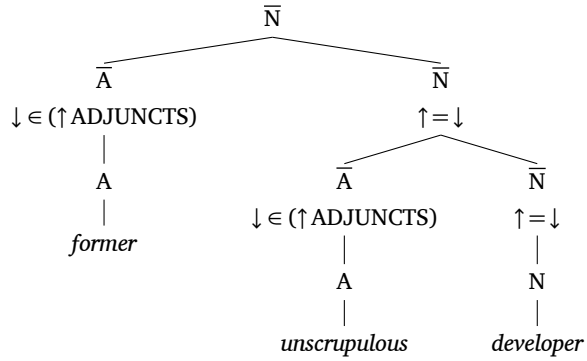
The problem for current LFG is that even if we adopt nested c-structures such as (13a) below, the f-structures will still be flat, because the $\overline{\text{N}}$s have to be introduced with $\uparrow = \downarrow$ annotations in order for the LFG analyses of agreement to work in examples such as *this/*these*

[5]There are also 'pseudo-modal' adjectives such as *fake*, which Partee (2010) analyses as being actually intersective, but exhibiting modal-like behaviour due to pragmatic accommodation effects.

[6]This example, which illustrates the essentially contextual nature of the phenomenon, is due ultimately to Georgia Green, and was pointed out to me by an anonymous reviewer.

former developer. So from a tree like (a) below, we still get the same form of structure as (10b), with the modifiers in an unstructured set that does not express the scope relations:

(13) a.



b.
$$\begin{bmatrix} \text{SPEC} & \text{INDEF} \\ \text{ADJUNCTS} & \left\{ \begin{bmatrix} \text{PRED} & \text{'former'} \end{bmatrix}, \begin{bmatrix} \text{PRED} & \text{'unscrupulous'} \end{bmatrix} \right\} \\ \text{PRED} & \text{'developer'} \end{bmatrix}$$

These structures could be interpreted using 'f-precedence' (Dalrymple 2001, 171–182), but Andrews (1983) shows that this introduces a problem: it is the order of concentricity out from the head that matters, rather than linear string order (as demonstrated by the behaviour of postnominal modifiers). For example, *a supposed American businessman* and *an American supposed businessman* are interpreted in the same way that the examples of (12) are, but (14) may be interpreted either way:[7]

(14) a supposed businessman from America

The interpretational problem is made concrete in the glue analysis of Dalrymple (2001, ch. 10), where the meaning-constructors for the modifiers will be able to operate on the two modifiers in either order,

[7] Sadler and Arnold (1994, p. 196) find that postnominal adjectives scope over prenominal ones, but they do not consider PPs, for which this does not appear to be the case, creating a problem for their interesting structural proposal. A possible account of the scope behaviour of postnominal APs is that that they are adjoined to DP in the manner argued for relative clauses by Vergnaud (1974) on the basis of examples such as *a man and a woman (who are) similar in their interests have a chance of getting along reasonably well*.

wrongly representing both sentences of (12) as ambiguous in the same way that (14) is.

To resolve this problem, I propose to use hybrid objects and distribution to support modification of the f-structures so as to follow the c-structure more closely. Then, glue or any other reasonable form of syntax-semantics interface can produce the correct interpretations without difficulties.

3.2 *Nesting structures*

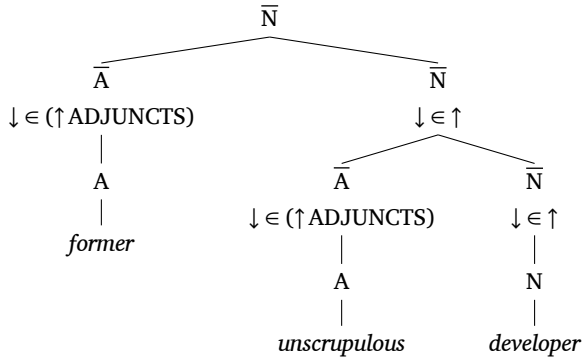
The f-structures I propose for the sentences of (12) are:

(15) a.
$$\begin{bmatrix} \text{ADJUNCTS} & \left\{\begin{bmatrix}\text{PRED} & \text{`former'}\end{bmatrix}\right\} \\ & \left\{\begin{bmatrix} \text{ADJUNCTS} & \left\{\begin{bmatrix}\text{PRED} & \text{`unscrupulous'}\end{bmatrix}\right\} \\ & \left\{\begin{bmatrix}\text{PRED} & \text{`developer'}\end{bmatrix}\right\} \end{bmatrix}\right\} \end{bmatrix}$$

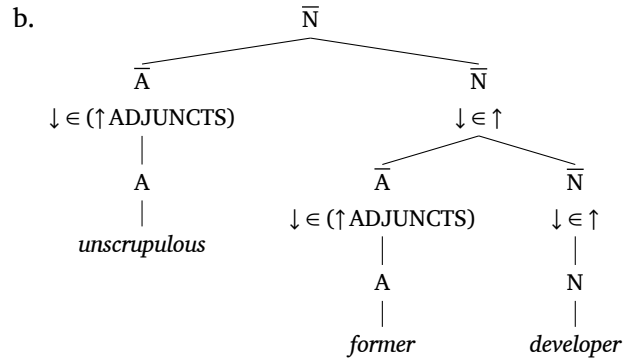
b.
$$\begin{bmatrix} \text{ADJUNCTS} & \left\{\begin{bmatrix}\text{PRED} & \text{`unscrupulous'}\end{bmatrix}\right\} \\ & \left\{\begin{bmatrix} \text{ADJUNCTS} & \left\{\begin{bmatrix}\text{PRED} & \text{`former'}\end{bmatrix}\right\} \\ & \left\{\begin{bmatrix}\text{PRED} & \text{`developer'}\end{bmatrix}\right\} \end{bmatrix}\right\} \end{bmatrix}$$

These use hybrid objects with singleton sets to preserve the information from the c-structure, and will be produced if the $\overline{\text{N}}$ expansions introducing the APs introduce their lower $\overline{\text{N}}$s with a $\downarrow \in \uparrow$ annotation rather than the usual $\uparrow = \downarrow$:

(16) a.



b.



For this to work, we need to assume that ADJUNCTS is non-distributive. To simplify the structures, we will also assume that PRED is non-distributive, but this assumption is not necessary and may be incorrect, as will be briefly discussed in the conclusion of this paper.

These structures provide a basis for semantic interpretation of these modifiers, which can be given with glue semantics, adapting the treatment of Dalrymple (2001). A brief description is provided in Appendix B, and we can explain the ambiguity of (14) in the obvious way by extending the phrase structure rules to expand $\overline{\text{N}}$ to $\overline{\text{N}}$ PP. These structures also account for other well-known properties of adjectival modification, such as that 'inner' adjectives cannot be ordered in front of intersective/subsective or modal adjectives:

(17) a. John is a tall/purported chemical engineer.
b. *John is a chemical tall/purported engineer.

There are further issues in adjective ordering to which the present proposals are relevant; but we turn instead to some phenomena of agreement and some issues concerning distribution.

3.3 *Coordination, agreement and undersharing*

As discussed in connection with example (7), Dalrymple and Kaplan assumed that the features of person, gender and number were nondistributive, because these features did not appear to be shared between the members of a coordinate structure and the whole. Subsequently, on the basis of previous work in HPSG and scholarship in various languages, especially Slavic ones, Wechsler and Zlatič (2000, 2003) made a strong case that agreement features should appear, often doubly, under under two sub-attributes, INDEX and CONCORD, the first primarily

involved in verb agreement, the second in concord within the NP. Person features seem to be restricted to INDEX, while gender and number are proposed to appear in both, usually with the same value, but sometimes different, in order to explain various agreement mismatches.

King and Dalrymple (2004) adapted and used these ideas to explain phenomena such as the apparent agreement anomaly in coordinations that are sometimes called 'close coordination', where there are two nominal phrases with different reference, but only one demonstrative that applies to both:

(18) a. This cat and dog are/*is friends.
b. *These cat and dog is/are friends.

They concluded that in English (specifically, as other languages differ), demonstrative pronouns show CONCORD agreement, which they proposed to be distributive, so that the demonstrative is singular, agreeing with the nominal heads of the two conjuncts. They further conclude that INDEX is nondistributive, and in this case is assigned on a semantic basis, so that the verb agreement is plural. Their structure (20) (p. 77) can be represented as (19) using our conventions:

(19)
$$\left[\begin{array}{ll} \text{SPEC} & \text{'this'} \\ \text{INDEX} & \left[\text{NUM}\ \ \text{PL}\right] \\ \text{CONCORD} & \left[\text{NUM}\ \ \text{SG}\right] \\ \left\{\left[\begin{array}{ll} \text{PRED} & \text{'boy'} \\ \text{INDEX} & \left[\text{NUM}\ \ \text{SG}\right] \\ \text{CONCORD} & [\ \ ] \end{array}\right], \left[\begin{array}{ll} \text{PRED} & \text{'girl'} \\ \text{INDEX} & \left[\text{NUM}\ \ \text{SG}\right] \\ \text{CONCORD} & [\ \ ] \end{array}\right]\right\} & \end{array}\right]$$

However, our proposed change in NP structure requires both kinds of features to be distributive. For as per Wechsler (2011), most nouns impose identity between the INDEX and CONCORD values of features, with equations such as (↑CONCORD NUM) = (↑INDEX NUM). This does not create a problem with the traditional LFG flat structures for NPs, but does with our present proposal, unless both attributes are distributive.

Were this not the case, the agreements wouldn't work in sentences like this:

(20) These alleged murderers are/*is surely guilty

The noun *murderers* would introduce a NUM PL feature and share it between INDEX and CONCORD; by the distributivity of CONCORD it would wind up on the demonstrative, but by the non-distributivity of INDEX it would not be passed up to the higher levels of the NP, and so singular agreement on the verb would be expected, instead of the plural that is actually required.

We therefore need to stipulate nondistributivity of INDEX in the close coordination construction. For this we propose to use the restriction notation from Kaplan and Wedekind (1993), in a rule like this:

(21) $\overline{\text{N}} \quad \rightarrow \quad \overline{\text{N}}^{+} \quad \text{Cnj} \quad \overline{\text{N}}$
$\qquad\qquad\qquad\quad \downarrow \in \uparrow/\text{INDEX} \qquad\quad \downarrow \in \uparrow/\text{INDEX}$

Consistently with its original use, the notation says that the f-structure of the upper $\overline{\text{N}}$ is the same as that of the daughters, except for the universally nondistributive attributes such as ADJUNCTS, and, in addition, the normally distributive INDEX attribute. Without such a stipulation, the distribution convention would cause the plural agreement of the verb to propagate into the conjuncts, and then be transmitted to their CONCORD-values and expressed morphologically. A similar undersharing specification is needed for the full NP/DP coordination rule, of which a preliminary version can be formed by replacing $\overline{\text{N}}$ with NP or DP in (21) above.

Such undersharing specifications are theoretically somewhat undesirable, but there is independent evidence that they are necessary. Dalrymple and Kaplan (2000, pp. 771–773) discuss the case of Xhosa, where the conjuncts of coordinated NPs have to agree in 'noun class' if anything agrees with them (but can disagree if nothing does):

(22) a. Umtwana uyagoduka.
(1/2GEND.SG)child (1/2GEND.SG)is going home
'The child is going home.'

b. umfana nomfazi
(1/2GEND.SG)young man (AND.1/2GEND.SG)woman
bayagoduka.
(1/2GEND.PL)are going home
'The young man and the woman are going home.'

c. *igqira nesanuse {a|zi-}yagoduka.
(5/6)doctor (AND.7/8)diviner {(5/6|7/8-}go home
trying to say: 'The doctor and the diviner went home.'

d. Igquira li-yagoduka nesanuse.
(5/6).doctor 5/6-is going home (AND.7/8)diviner
'The doctor is going home with the diviner.'

e. Isanuse si-yagoduka niguireanumber
(7/8).diviner 7/8-is going home (AND.5/6).doctor
'The diviner is going home with the doctor.'

Sentence (a) illustrates agreement with a singular, noncoordinated noun; (b) with a coordinated noun where the conjuncts have the same gender; (c) the failure of such a case where the genders differ; (d,e) an alternate construction that can be used when the 'classes' differ.

Their proposal is that these examples involve a distributive attribute 'class' rather than nondistributive gender, but distributivity appears to be the only respect in which 'class' is clearly different from gender. Indeed, in his discussion of the Bantu 'class' system, Corbett (1991, pp. 43–46) notes that in early Bantu work, 'class' referred to the combinations of a kind of gender with number, so that 'animate' singular was class 1, animate plural class 2, etc. But this view accords too little recognition to the regular relation between the semantically singular and plural classes, which indicates that the gender-like property should be dissociated from number, which is further supported by examples like (b) above, where two class 1 nouns trigger agreement by a class 2 prefix.

Corbett thereby distinguishes gender from number, and designates such postulated genders as '1/2' and '3/4', based on the original class terminology. This notation maintains a convenient and useful amount of contact with the earlier tradition, while providing more satisfactory analyses. Corbett calls these categories genders, and their only apparent difference from familiar traditional genders is their different behaviour with respect to distribution. Since the traditional Bantu class pairs seem to show no major differences besides behaviour under distribution from other putative genders, there is no basis for treating them as a different kind of attribute.[8] Therefore distribution

[8]Another possible difference, pointed out by an anonymous referee, is that gender is subject to resolution and class is not. But resolution is an extremely

is not a sufficient basis for distinguishing Bantu 'slashed classes' from other instances of gender.

Instead, I suggest that gender is normally undershared in coordinate structures, presumably for the functional reason that this allows a wider range of coordinations to be generated. However, such an undersharing stipulation happens to be absent from Xhosa (the availability of a semantically approximately equivalent comitative construction might be a relevant factor). This treatment is better motivated if we can find other kinds of situations that can be well-analyed as stipulated undersharing, to which we turn in the next subsection.

3.4 *Agreement discontinuities*

Pesetsky (2013), Ouwayda (2014), Landau (2016) and Puškar (2017) discussed another kind of phenomenon that can be analysed in terms of stipulated nondistributivity involving singleton sets. The treatment here is brief, due to the number of languages involved that don't seem to have much in the way of relevant previous work in LFG, but the phenomena are striking.

The basic phenomenon is that either gender or number agreement within an NP shifts from grammatical (as determined by the head) to semantic. Sentence (a) below is a Russian example involving case, while sentence (b) is a Modern Hebrew example involving number:

(23) a. U nas byl-a očen xoroš-aja zubn-oi vrač-ъ.
of us was-FEM very good-FEM dental-MASC doctor-MASC
'We had a very good female dentist.' (Pesetsky (2013, p. 38), citing earlier work)

b. ha-be'alim ha-pratiyim ha-axaron šel ha-tmuna
the-owner(PL) the-private(PL) the-last(SG) Pos the-painting
haya …
was(SG)
'The last private owner of the painting was [the psychoanalyst Jacques Lacan].' (Landau (2016, p. 1005); naturally occurring example from Wikipedia)

complex phenomenon, to the extent that one can actually doubt whether it really exists as a concept of grammatical theory, and our knowledge of the Bantu languages with noun class is relatively limited. Therefore, I do not find this to be a clear difference.

The background to (a) is that in Russian, professional nouns are invariably masculine in their grammatical gender, as shown by the masculine agreement of the adjective *zubnoj*, but if the referent is female, the gender has the possibility of switching (it can also stay masculine, or switch at various places). In (b), the background is that the word *be'alim* in Hebrew is grammatically plural but can have singular reference, but if the reference is singular, adjectives and the main predicate can switch to singular. A significant commonality between both examples is that if a switch occurs overtly in the nominal, the verb must follow suit, and, within the nominal, the switch must obey the concentricity hierarchy: if a more inner element switches, all the more outer ones must switch too, with opposite linear order in the two languages. This also happens with the other case of agreement discontinuities discussed by both Landau and Puškar: gender (class) agreement in Chichewa.[9]

An initial thought might be that we could use INDEX and CONCORD to analyse this, and indeed Landau provides such an analysis within the Minimalist Program. But given the flat structures of current LFG, INDEX and CONCORD don't help, because both attributes will be attributes of the same f-structure. Therefore, if they are equated or non-equated anywhere in that structure, they will be so equated or non-equated everywhere, providing no basis for explaining concentricity.

We can do better with nesting of singleton sets and undersharing. First, a note on 'grammatical' versus 'semantic' agreement: cross-linguistically, agreeing modifiers will almost always show 'grammatical' agreement if they are modifying something with grammatical gender or number (*pluralia tantum*), but will show 'semantic' agreement if there is no overt grammatical agreement trigger, as seen in these examples from Modern Greek:

(24) a. I arsenikí arákhni huntsman fénete na méni
the(F) male(F) spider(F) huntsman seems to remain
akíniti ke eksouthenoméni.
motionless(F) and exhausted(F)
'The male huntsman spider seems to remain motionless and exhausted.'[10]

[9]These concentricity effects are currently treated in the Minimalist Program as an aspect of the 'Agreement Hierarchy' of Corbett (1979).

[10]`http://www.inewsgr.com/122/apokosmo-vinteo-me-trichoto-kai-`

b. Íme étimi.
I am ready(F)
'I am ready (female speaking, not male).'

The LFG + glue literature does not provide an explicit account of how semantic agreement works/integrates with syntactic agreement; the nearest approach being Wechsler (2011) in a non-glue LFG formulation. The following, based on Wechsler, seems workable:[11]

(25) a. Grammatical gender and number associated with nouns are introduced by defining equations on the lexical entries of those nouns, without meaning-constructors specific to the features.

b. Semantically transparent gender and number associated with nouns are introduced on those nouns by defining equations with associated meaning-constructors.

c. Agreeing items all have free choice between:

i) introducing a constraining equation with no meaning constructor (grammatical agreement),

ii) introducing a defining equation with a semantically appropriate meaning-constructor (semantic agreement).

For work relevant to the distinction between (a) and (b) in Greek, see Merchant (2014) and Alexiadou (2017). Rule (c) implies that lexical entries of agreeing items such as *étimi* 'ready'(Fem.Nom.Sg) all have disjunctive specifications; this is notationally a bit awkward but can be done with 'templates' (a kind of macro used in LFG, as briefly discussed below), and is similar to the 'Agreement Marking Principle' of Wechsler (2011, p. 1009).

As exemplification of the proposed principles, in (24b), the adjective *étimi* 'ready' would have a defining equation and a feminine gender meaning-constructor; whereas in (24a), the noun *arákhni* 'spider' would have a defining equation for feminine gender without any associated meaning-constructor, while the other adjectives would have constraining equations for gender, once again without meaning-constructors.

tromaktiko-plasma-prokalei-anatrichila-sto-internet.htm; viewed Jan 12, 2018.

[11] Note that the notationally complex disjunction in (c) can be managed with templates.

To get the Russian agreement discontinuity, we use an alternate expansion of NP that undershares GENDER in CONCORD and INDEX (unfortunately, we need to do both). There is a further restriction: these discontinuities can only happen in the nominative case, leading to the following rule:

(26) NP → NP
(↑CASE) = NOM
↓ ∈ ↑/(CONCORD|INDEX) GEND

The use of the typically disjunctive '|' symbol is motivated by the consideration that a gender feature is not distributed if it lies in either the INDEX or the CONCORD bundle. We need to do this in order to change both the presumably INDEX agreement on a main verbal predicate such as *byl-a* 'was-F' and an adjectival one such as *xoroš-aja* 'good-F' in (23a). The rule (26) only has a discernable effect in singular NPs because the genders are neutralized in the plural. When (26) applies, any higher agreeing items will have to have their gender features interpreted semantically. Another, technical, point is that for (26) not to run afoul of the offline parsability constraint (Kaplan and Bresnan 1982, p. 266), we need to adapt the constraint so as to allow a node of type X to dominate another node of type X as long as they introduce different annotations.

3.5 *Conclusion*

We have applied hybrid objects with singleton sets to adjectival modification constructions, proposing a solution to issues that have remained largely unsolved in LFG. A further, general observation is that per conventional LFG + glue, we should expect that the linear or hierarchical arrangement of modifiers would normally impose no solid restriction on interpretation, in a way comparable to what we often find with quantifier scope. As far as I am aware, this is extremely rare or nonexistent with modifiers, and the sensitivity of scope to concentric arrangement extends to somewhat exotic constructions such as the Modern Greek 'polydefinite' construction (Velegrakis 2011, esp. pp. 31–35).

Nordlinger and Sadler (2008) and Sadler and Nordlinger (2010) proposed applying sets to NP structure, in Australian languages. They do not use singleton sets, but do have problems with making distribution work; the undersharing mechanism proposed here could help.

4 COMPLEX PREDICATES IN ROMANCE

Romance complex ('restructuring') predicates pose a classic problem. On the one hand, they are 'monoclausal', as evidenced by clitic climbing and other phenomena that seem to show that they constitute a single clause. On the other hand, they demonstrate 'respect for the tree': both their interpretation and the distribution of their verbal markers appear to depend on the tree structure,[12] both of which are problematic for LFG, which assumes that both verbs inhabit a single clause in f-structure. These points are illustrated by these examples from Catalan (Alsina p.c.), repeated from Andrews (2007):

(27) a. L' acabo de fer llegir al nen.
it I.finish of make.INF read.INF to the boy
'I just made/I finish making the boy read it.'

b. La faig acabar de llegir al nen.
it.F I.make finish.INF of read.INF to the boy
'I make the boy finish reading it (say, a map ([GND FEM])).'

Here, the final verb is generally considered to be the 'main' verb, whereas the (two) preceding ones would be considered 'light' verbs.

The appearance of clitics *L'* (gender-ambiguous) and *La* expressing an argument of the main verb on the first light verb provides one of the arguments that the construction is monoclausal. The other is that the arrays of the arguments of the individual verbs appear to be combined into one, which obeys the rules for the array of grammatical relations for transitive and ditransitive predicates. In particular, *the boy*, the Agent and expected subject of the Caused verb, is expressed as an *a*-object, the normal grammatical relation for the Recipient of a ditransitive, and there is only one bare NP object, as occurs regularly in the Romance languages.

Various other languages combine indications of, on the one hand, hierarchical embedding of the structure headed by the Caused verb within one headed by the Causer verb, and, on the other, fusion of the two levels of the structure into something that appears for at least some purposes to be a single clause. An important example in the LFG

[12] The linear order is another possibility, but this doesn't seem to be workable, and there would be no explanation for why the relevant linear order is reversed for Hindi/Urdu, as discussed below.

literature has been Hindi/Urdu (Butt 1995), most recently analysed within LFG + glue by Lowe (2015). He proposes (p. 442) the f-structure of (28b) for the example (28a):

(28) a. Amu-ne bacce-se haathii pinc
Amu-ERG child.OBL-INSTR elephant pinch
kaar-vaa-yaa.
do-CAUSE-PERF.MSG
'Amu caused the child to pinch the elephant.'

b. $\begin{bmatrix} \text{PRED} & \text{'pinch'} \\ \text{CAUSE} & + \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'Amu'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'elephant'} \end{bmatrix} \\ \text{OBJ}_\theta & \begin{bmatrix} \text{PRED} & \text{'child} \end{bmatrix} \end{bmatrix}$

The PRED-value is the main verb; the causative verb is represented as a non-PRED feature value; and glue semantics is used to get the grammatical relations correctly associated with their semantic roles. This single-layer f-structure analysis, which we could describe as 'fully monoclausal' due to having only a single layer of f-structure like an ordinary simple clause, works reasonably well for Hindi/Urdu; whereas in Romance languages, such an analysis is more problematic, as we discuss in the next section.

4.1 *Problems with the fully monoclausal analysis*

There are three problems: the determination of forms, the multiplicity of light verbs, and the relevance of order. We consider each in turn.

In Romance languages[13] an infinitive may appear with or without an additional verb marker such as *a* or *de*, while some verbs instead take a present or past participle without any additional marker. Taking as examples (27) and (33), we find the following form determinations:

(29) a. *acabar* 'finish' is followed by *de* + infinitive
b. *fer* 'cause/make' is followed by a bare infinitive
c. *poder* 'can' is followed by a bare infinitive

[13]There is a form-determination problem in Urdu, discussed later; it is much more limited than in Romance, and does not provide as much difficulty for LFG.

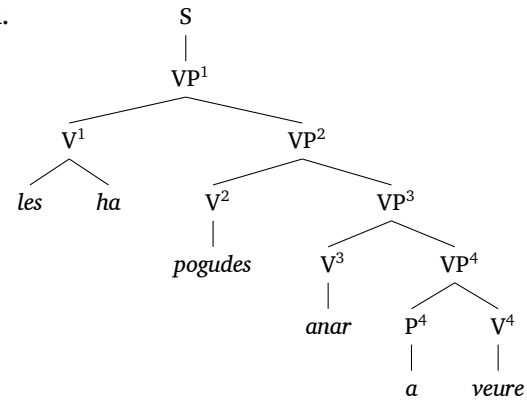
d. *haver* 'perfect auxiliary' is followed by a past participle

e. *anar* 'go to' is followed by *a* + infinitive

Solà (2002) provides many more examples. The original solution to this problem was to add an additional projection called m-structure (Butt *et al.* 1996, Butt *et al.* 1999). This can be made to work, but has not fared very well, as we now discuss.

M-structure was originally proposed to be a projection directly from c-structure, and could be thought of as a kind of enrichment of the c-structure that includes certain inflectional features, in particular the ones that light verbs impose on their 'semantic complements', which follow them in Romance languages. The lexical entries of verbs would put their verbal form and marker features on m-structure, and the c-structure rules would specify the m-structure of a VP complement as the 'DEP'-value of the m-structure of its containing VP. Light verbs would furthermore specify what features their DEP-values should contain. Example (33c) below would then have the following c- and m- structures, where the correspondence is indicated by numerical superscripts rather than dotted lines in order to reduce clutter:

(30) a.



b.

$$\begin{bmatrix} \text{FIN} & + & & & & & \\ 1 & \begin{bmatrix} \text{DEP} & 2 & \begin{bmatrix} \text{VFORM} & \text{PERFP} & & & \\ \text{DEP} & 3 & \begin{bmatrix} \text{VFORM} & \text{INF} & & \\ \text{DEP} & 4 & \begin{bmatrix} \text{VFORM} & \text{INF} \\ \text{VMARK} & \text{A} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Such structures are provided by appropriate placement of annotations like these in the phrase-structure rules, where '$*$' means the c-structure node the annotation appears on, '$\hat{*}$' the mother of that node, and $_{\mathrm{m}}$ the m-structure of the node referred to:

(31) a. $(\hat{*}_{\mathrm{m}}\ \mathrm{DEP}) = *_{\mathrm{m}}$

b. $\hat{*}_{\mathrm{m}} = *_{\mathrm{m}}$

Since the m-structure comes off c-structure rather than f-structure, it is not a problem if the f-structure is flat. Note also that the clitic pronoun *les* does not appear in the m-structure, because m-structure is not a full representation of the hierarchical structure of a sentence, and, in particular, does not include the grammatical relations. If the clitic did have an m-structure, it would be disconnected from that of the verbs.

M-structure does what it is supposed to do, but comes at a certain cost. First, we have an entire additional projection for which relatively few additional uses have been proposed, and for which there is no motivation whatsoever in many languages, including richly inflected ones such as Greek or Icelandic (their causatives are either fully morphological or unambiguously biclausal). Indeed, this projection now perhaps has no current uses at all in its original form, as an independent projection from c-structure. For example, Belayev (2013) applies a concept of m-structure to person agreement in the East Caucasian language Dargwa, but he uses the proposal of Frank and Zaenen (2004) that m-structure comes off f-structure rather than c-structure. Frank and Zaenen manage to make this proposal work for French, where there is reasonable evidence that the light verbs are introduced in a verbal cluster that does not include any verbal complements, and they are restricted to a small number of auxiliaries. But it is very hard to imagine how their proposal could extend to southern Romance languages, where not only is the VP-complement 'right-branching structure' well argued for and generally accepted (e.g. Manning 1996, Alsina 1997), but also, the inventory of light verbs is much larger, and not confined to any class that could reasonably be described as 'auxiliaries'.

This leads to our second problem. Solà (2002, pp. 227–229) gives a substantial but not complete collection of restructuring verbs. In addition to various aspectual concepts and the verbs 'come' and 'go',

the collection contains 'learn', 'go up' (to do something) and 'pass by' (to do something), yielding examples such as these:

(32) a. Ho he après a fer.
it I have learned to do.INF
'I have learned to do it.'

b. El pasaré a saludar.
him I will pass by to greet
'I'll pass by to greet him.'

c. L' he baixat a buscar.
him/her I have gone down to fetch.INF
'I have gone down to fetch him/her.'

Solà cites them as evidence against Cinque's proposal to treat light verbs as heads of functional projections, on the basis that they have too much lexical content to plausibly serve in this way. But their lexical richness is even more problematic for the featural representation of example (28).

The flatness of the featural representation also fails to account for 'respect for the tree', for which we have not only Alsina's examples above, but some additional ones from Solà (2002, p. 238):

(33) a. Les pot aver vistes.
them.F can.3SG have.INF see.PSTPART.FPL
'He/She can have seen them(F).'

b. Les ha pogudes veure.
them.F have.3SG can.PSTPART.FPL see.INF
'He/She has been able to see them(F).'

c. Les ha pogudes anar a veure.
them.F have.3SG can.PSTPART.FPL go.INF to see.INF
'He/She has been able to go to see them(F).'

Even if we accept the idea of representing each item with a feature, there is still the problem of getting the interpretation correctly determined by the order.

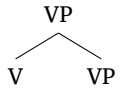
One could think of trying to do something with the notion of 'f-precedence', but, as far as I can work out, glue semantics does not include any way of saying something like 'if you are my semantic argument, I must precede you' in a situation where the structure is flat,

and all items have the same f-structure and therefore s-structure. Andrews (2007) makes a proposal for a general principle, but it involved some additions to the theory, and did not get general uptake by the LFG community. Furthermore, it does not appear to be applicable to the problems with adjectival modifiers discussed in the previous section. But I claim that singleton sets with undersharing can solve all of these problems.

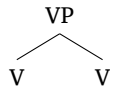
4.2 *A solution with hybrid objects and undersharing*

The proposal is that light verbs are introduced in the structures in (34) below: structure (a) applies when a right-branching VP seems indicated (Catalan and Spanish), whereas (b) applies when the verbs seem to form a cluster (French).

(34) a.



b.

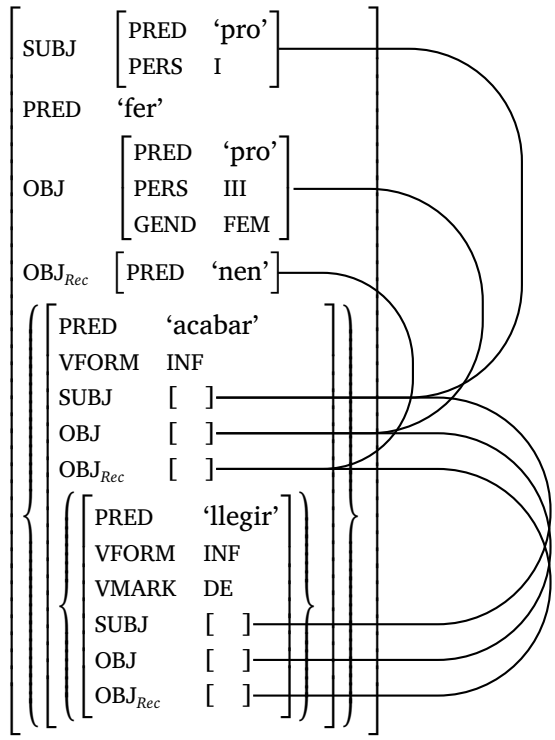


The orders are expected to be reversed in verb final languages, unless diachronic changes have occurred and made the rules more complex. Hindi/Urdu is an example with verb-final order and both (a) and (b) structures, but with the order of the daughters reversed (Butt 1995).

Superficially similar structures that do not in fact appear to involve clause-union can have the same c-structure form, but with the lower VP introduced as value of XCOMP, OBJ, or whatever else seems appropriate on the basis of the relevant evidence.

Now, example (27b) will get an f-structure like (35), with the grammatical relations shown as shared through all the levels. But we don't try here to represent the lexical specifications of the predicates for their arguments, because this involves issues of linking theory that we take up below:

(35)



Form-determination can then be accomplished via the f-structure by specifications like these:[14]

(36) a. *acabar*: (↑∈ VFORM) = INF, (↑∈ VMARK) = DE.

b. *haber*: (↑∈ VFORM) = PASTPART, ¬(↑∈ VMARK).

c. *fer*: (↑∈ VFORM) = INF, ¬(↑∈ VMARK).

The proposed structure therefore solves both of the problems discussed at the beginning of this section, with the provision that we need to treat the VFORM and VMARK as nondistributive. Nonetheless, they behave distributively in coordination, requiring undersharing in complex predicates, as we discuss in the next subsection.

[14] Note that they are technically functionally uncertain, due to the membership relation; notwithstanding, this is moot because the set is a singleton.

4.3 *Distributive issues*

We see in the following examples obligatory distribution of the infinitive VFORM and possibly optional distribution of VMARK:

(37) a. acabà de riure i (de) plorar.
finish.PRET.3SG VM laugh.INF and (VM) weep.INF
'He/she stopped laughing and crying.'

b. Quan acabis de llegir l'article i (de)
when finish.SUBJ.2SG VM read.INF the-article and (VM)
fer-ne el resum, avisa'm.
make-of it the summary, advise-me.
'When you finish reading the article and summarizing it, let me know.'
(Alsina p.c.)

Although both versions of (b) are acceptable, the one with the second *de* included is more formal, to the extent that, if omitted, it might be supplied by a copy editor (Alsina p.c.). We can account for this with two assumptions: first, that there is no undersharing of VMARK in coordinate structures; and second, that the verbal marker is introduced in a slightly higher projection than VP, either the higher or the lower able to be conjoined. Formal style prefers coordinating the higher one.

Distribution of infinitive, gerund and past participle VFORM in coordinate structures is illustrated here:

(38) a. La Maria fa riure i plorar el nen.
the Mary makes laugh and cry the boy
'Mary makes the boy laugh and cry.'
Alsina (1997, p. 222)

b. La Maria està rient i plorant.
the Mary is laughing and crying
(Alsina p.c.)

c. La Maria ha rigut i plorat.
the Mary has laughed and cried
(Alsina p.c.)

We can resolve the non-distribution issue by stipulating undersharing in the light verb VP rule, which can now be formulated as:

(39) VP → V VP
↓ ∈ ↑/VFORM /VMARK

These constructions were originally complement structures, which explains the undersharing stipulations, since features are not normally shared between complements and their heads.

Adverb placement constitutes a potential problem for the present treatment of distribution. The previous section and the discussion of frequency adverbs in Andrews (1983) indicate that the ADJUNCTS attribute is not distributive. However, Andrews and Manning (1999, p. 55) offer a contrary example:

(40) a. He fet beure el vi a contracor a la Maria.
I have made drink the wine against *x*'s will to the Mary
'I have made Mary drink the wine against her/my will.'

b. Volia tastar amb molt d'interès la cuina tailandesa.
I wanted to taste with much interest the cuisine Thai
'I wanted to taste Thai food with much interest.'
(*with much interest* most naturally modifying *want*)

Catalan has the possibility of putting the object NP after the verb in simple clauses as well as restructuring ones. The two examples below are both fine without any obvious intonational peculiarities (Alex Alsina, p.c.), although the traditional doctrine is that the NP would normally go first:

(41) a. entendràs les meves raons de seguida.
understand.FUT.2SG the my reasons right away
'You'll understand my reasons right away.'

b. entendràs de seguida les meves raons.
'You'll understand my reasons right away.'

Further examples with the NP after an adverbial PP can be found on the web:

(42) a. Llegiré amb calma tota la teva disertació.
read.FUT.1SG with calmness all the your 'dissertation'
'I will read with calmness your entire 'dissertation'.'[15]

[15] http://hemeroteca.e-noticies.com/edicio-1168/popups/popVerComentariosElemento_asp_idSeccion_3_idSubSeccion__id_2000633.htm; viewed 16 Feb 2018

b. Llegeixo amb atenció el teu post.
read.1SG with attention the your post
'I read (present tense) your post with attention.'[16]

Therefore, there is clearly a position for NPs at the end of the VP, after an adjunct PP. Also, since the OBJ grammatical relations are distributive, an NP can appear after an adjunct PP in the upper VP, while still functioning as the object of the lower verb. We can therefore explain the examples of (40) without having ADJUNCTS be distributive.

4.4 *Linking theory*

We now have almost everything we need except for a linking theory to account for the facts of subcategorization. There are a considerable number of options to choose from in the literature on these constructions, including those of Alsina (1996), Andrews and Manning (1999), and Andrews (2007). But here I will do something different, and propose an account of linking on the basis of the 'Kibort-Findlay Mapping Theory', henceforth KFMT, although I won't attempt a full integration of the analysis with that theory. KFMT is the development of the mapping theory of Kibort (2013) by Findlay (2016), also used in Asudeh *et al.* (2014).[17] Its drawback for our purposes is that it has not yet been adapted to the demands of Romance languages, which show some differences from the Germanic and Bantu languages that most LFG lexical mapping theories other than Alsina's appear to be focused on. The reason for developing KFMT is that, unlike its predecessors, it is both fully within the formal theory of LFG, and capable of handling clause-union constructions.

The key to this capability is that it makes heavy use of glue semantics, in a way that allows it to deal in a straightforward way with the problem of suppressing the linking of the Agent argument of the Caused-verb to a subject grammatical function. Classic LMT works on a predicate-by-predicate basis, supplying grammatical relations to underspecified argument positions, which makes subject-suppression in complex predicate constructions difficult to achieve if they are viewed as actually having two predicates, while the proposals noted above, of

[16] `http://interaccio.diba.cat/blogs/2015/intent-dapuntar-pros-contres-gestio-comunitaria-cultura`; viewed 16 Feb 2018

[17] I am indebted to Ash Asudeh for suggesting that I try this.

Alsina on one hand and Andrews and Manning on the other, try to address this problem with devices that are not clearly and fully within the LFG formalism.

KFMT terminologically abandons the popular idea of 'argument structure', but replaces it with an elaboration of the 'semantic projection' of glue semantics. This is similar enough to argument structure that perhaps the concepts are being fused, rather than one replacing the other. The semantic projection is a projection from f-structure, and the novelty is to populate it with attributes such as ARG_1, ARG_2 and more, which reflect a classification of semantic roles in terms of their typical syntactic behaviour.

ARG_1 is like the 'external argument' of GB/Minimalism, the 'I' of relational grammar, or the 'Actor' of Role and Reference Grammar, while ARG_2 is like the non-oblique 'internal argument' of GB and Minimalism, the 'II' of Relational Grammar, or the 'Undergoer' of Role and Reference Grammar. ARG_4 and above are obliques, while ARG_3 is complicated, and will be discussed shortly. KFMT also uses Davidsonian event semantics, with an event variable. The meaning-constructor for a transitive verb such as *llegir* 'read' would be:

(43) $\lambda y x e.\mathit{Llegir}(e) \wedge \mathit{Agent}(x,e) \wedge \mathit{Patient}(y,e)$:
$(\uparrow_\sigma \text{ARG}_2) \multimap (\uparrow_\sigma \text{ARG}_1) \multimap (\uparrow_\sigma \text{EV}) \multimap \uparrow_\sigma$

If this is added to a lexical entry that introduces the PRED-value 'llegir' into the f-structure, then we get the following pieces of f- and s-structure connected by the semantic projection σ as the solution (the λ-term for the meaning not yet included):

(44) $\left[\text{PRED} \quad \text{'llegir'}\right] \cdots\cdots_\sigma \cdots\cdots\blacktriangleright \begin{bmatrix} \text{EV} & [\ ] \\ \text{ARG}_1 & [\ ] \\ \text{ARG}_2 & [\ ] \end{bmatrix}$

The '$\uparrow_\sigma$' at the end of (43) will associate the output of the meaning-constructor with the semantic projection of the f-structure in (44), but we need some additional machinery to associate the ARG_i-values there with the grammatical relations that will express the arguments.

This is accomplished by the linking theory, which provides specifications of equations that equate the semantic projection of the bearer of a grammatical function with an ARG_i value. These specifications are

highly compressed by templates.[18] A relatively simple one is the template @ARG2, which is an abbreviation for instructions to optionally add the following specification to a lexical entry:

(45) $<(\uparrow\{\text{SUBJ}|\text{OBJ}\})_\sigma = (\uparrow_\sigma \text{ARG}_2)>$

In addition to the optionality of the whole equation as indicated by the angle brackets, there is an optional choice notated by the | within the equation, which allows for the object-to-subject 'promotion' that is a characteristic of the passive. The optionality of the equation allows for the NP argument to fail to be realized in f-structure, as long some other component of the lexical entry will provide a suitable meaning to the glue-semantics, as discussed by Asudeh *et al.* (2014). If this does not happen, then the glue assembly will fail due to resource deficiency.

A slightly more difficult example is the ARG_1 specification, which expands to this:

(46) $<(\uparrow\{\text{SUBJ}|\text{OBL}_\theta\})_\sigma = (\uparrow_\sigma \text{ARG}_1)>$

Here, OBL_θ allows for the expression of an ARG_1 as a prepositional phrase in the passive, with some additional facilities, not discussed here, optionally supplying this argument in the glue semantics if there is no *by*-object in f-structure.

We will need a third kind of specification for the *a*-objects of Romance languages, which don't exactly fit into any of the categories developed in KFMT so far. I suggest that they are a variety of ARG_3, which are generally taken to be objects that can alternate between OBJ and OBJ_θ. Romance languages don't appear to have evidence for any such alternation, at least at the level of overt form,[19] so that in these languages, I suggest that ARG_3 are the *a*-objects, which are how Romance languages spell out OBJ_θ. This gives us @ARG3 as abbreviating this specification:

(47) $<(\uparrow\text{OBJ}_\theta)_\sigma = (\uparrow_\sigma \text{ARG}_3)>$

[18] A form of macro originally part of XLE implementation of LFG, but recently being explored more aggressively as an abbreviatory device for the linguistic theory.

[19] There are subtle arguments from the Minimalist Program that such alternations exist in languages where they are not morphosyntactically obvious, for example Anagnostopoulou (2003, pp. 230–234) on *a*-objects in Spanish.

Then, by virtue of other parts of the grammar, OBJ_θ is always realized as an *a*-object. With this background, we can consider the linking with restructuring predicates.

With intransitive light verbs such as aspectuals, the light verb has no effect on the available arguments; by extension, any theory that works for non-restructuring constructions will work for intransitive light verbs. But with the causatives we have the troublesome phenomenon of the Causee Agent being expressed as an object if the Caused verb is intransitive, but an *a*-object if it is intransitive:

(48) a. L' elefant fa riure les hienes.
the elephant makes laugh.INF the hyenas
'The elephant makes the hyenas laugh.'

b. Els pagesos fan escriure un poema al follet.
the peasants make write.INF a poem *a*.the elf
'The peasants make the elf write a poem.'

Furthermore, there is evidence that the Causee Agent is never in any way associated with the SUBJ-grammatical function, as discussed by Andrews (2007), who in turn further developed the arguments of Alsina (1996). So we need to completely suppress any possible linking of it to a SUBJ grammatical function.

The formal apparatus of KFMT allows us to do this by implementing an s-structure version of the glue semantics analysis provided in Asudeh (2005) of functional control by an argument of a higher verb.[20] The idea is that if a predicate calls for an argument of type $e \multimap t$, then any argument of that type which this applies to cannot accept any additional argument associated with the e, since this would cause 'resource surplus' in the glue semantics.

Therefore, the widely accepted 'three place causative' predicate can have a meaning-constructor like this:

(49) $\lambda P\, y\, x\, e.\mathit{Cause}(e) \wedge \mathit{Agent}(x, e) \wedge \mathit{Causee}(y, e)$
$\wedge\, (\exists d)(\mathit{Caused\_Event}(d, e) \wedge P(y)(d))$:
$[(\uparrow \in \sigma\ \text{ARG}_1) \multimap (\uparrow \in \sigma\ \text{EV}) \multimap (\uparrow \in \sigma)] \multimap$
$(\uparrow_\sigma \text{ARG}_{\{2|3\}}) \multimap (\uparrow_\sigma \text{ARG}_1) \multimap (\uparrow_\sigma \text{EV}) \multimap \uparrow_\sigma$

[20] Lowe (2015) also accomplishes complete subject suppression in a different way, which does not appear to be compatible with the present syntactic analysis, although it also employs KFMT.

The first two lines represent the meaning, in Davidsonian event semantics, while the third line is the glue term for the VP 'Caused' argument, with open positions for the ARG_1 and the event variable. The remaining arguments and the return of type t appear on the final line. The first argument on this line can be either an ARG_2 or an ARG_3; this will be discussed below.

A typical constructor for a verb that this would apply to would be (43), repeated below for convenience:

(43) $\lambda yxe.\mathit{Llegir}(e) \wedge \mathit{Agent}(x,e) \wedge \mathit{Patient}(y,e)$:
$(\uparrow_\sigma \text{ARG}_2) \multimap (\uparrow_\sigma \text{ARG}_1) \multimap (\uparrow_\sigma \text{EV}) \multimap \uparrow_\sigma$

If these are introduced in combination with the f-structure and s-structure of (50), their instantiated result would be (51), where labels are used to connect the semantic projection and glue literals:

(50)
$$\left[\begin{array}{ll} \text{PRED} & \text{'fer'} \\ \text{SUBJ} & [\ ] \\ \text{OBJ} & [\ ] \\ \text{OBJ}_\theta & [\ ] \\ \left\{\left[\text{PRED} \quad \text{'llegir'}\right] \ldots\right\} & \end{array}\right] \ldots\!\!\rightarrow \left[\begin{array}{ll} \text{EV} & [\ ]^a \\ \text{ARG}_1 & [\ ]^b \\ \text{ARG}_3 & [\ ]^c \end{array}\right]^g$$
$$\left[\begin{array}{ll} \text{PRED} & \text{'fer'} \\ \text{SUBJ} & [\ ] \\ \text{OBJ} & [\ ] \\ \text{OBJ}_\theta & [\ ] \\ \left\{\left[\text{PRED} \quad \text{'llegir'}\right] \ldots\right\} & \end{array}\right] \ldots\!\!\rightarrow \left[\begin{array}{ll} \text{EV} & [\ ]^d \\ \text{ARG}_1 & [\ ]^e \\ \text{ARG}_2 & [\ ]^f \end{array}\right]^h$$

(51) a. $\lambda Pyxe.\mathit{Cause}(e) \wedge \mathit{Agent}(x,e) \wedge \mathit{Causee}(y,e)$
$\wedge (\exists d)(\mathit{Caused\_Event}(d,e) \wedge P(y)(d))$:
$(e \multimap d \multimap h) \multimap c \multimap b \multimap a \multimap g$

b. $\lambda yxe.\mathit{Llegir}(e) \wedge \mathit{Agent}(x,e) \wedge \mathit{Patient}(y,e)$:
$f \multimap e \multimap d \multimap h$

Before we can apply (a) to (b) with implication elimination, we have to satisfy the first argument (label f) of (b), either by applying it to a 'real' argument such as perhaps *War and Peace*, or to a 'dummy' argument supplied as an assumption for later implication introduction; we'll represent the result of this with a **w** substituted for y:

(52) $\lambda xe.\mathit{Llegir}(e) \wedge \mathit{Agent}(x,e) \wedge \mathit{Patient}(\mathbf{w},e)$:
$e \multimap d \multimap h$

Now if we apply (51a) to (52) with implication elimination, we get the following after β-reduction:

(53) $\lambda y\, x\, e.\mathit{Cause}(e) \wedge \mathit{Agent}(x,e) \wedge \mathit{Causee}(y,e) \wedge$
$(\exists d)(\mathit{Caused\_Event}(d,e) \wedge \mathit{Llegir}(d) \wedge \mathit{Agent}(y,d) \wedge$
$\mathit{Patient}(\mathbf{w},d))$: $c \multimap b \multimap a \multimap g$

The application of the causative verb to the Caused one is specified in terms of the s-structure and the $\in$ relationship in f-structure, and therefore can proceed without linking, but the NP arguments require this, to which we now turn.

In KFMT, the linking equations are optional, with the result that the Causee ARG_1 doesn't have to be linked to anything, which is good, because if it does try to link, this will cause assembly failure due to resource surplus. But the remaining ones either must link, or require some other meaning-constructor to match them up with something in meaning-assembly, as discussed by Asudeh *et al.* (2014).

Since this is an active sentence, there is no alternative to linking the Causer ARG_1 with a syntactically represented argument. Therefore, the Causer Agent/ARG_1 must be a SUBJ, so the remaining ARG_is must be apportioned between OBJ and OBJ_θ. If the caused verb is transitive, it will have an ARG_2, whose only options are SUBJ and OBJ; the former is already taken, so it must get linked to OBJ. The Causer Object, on the other hand, will have to take its ARG_3 option (as notated in (49)), and be realized as OBJ_θ. With an intransitive Caused verb, we encounter a problem, which is that constructor (49) provides two possibilities for its 'Caused' argument, ARG_2 and ARG_3, but only the former is possible. This requires a stipulation, which can be a (constraining) implication saying that if there is an OBJ_θ, there must be an OBJ:

(54) $(\uparrow \text{OBJ}_\theta) \supset (\uparrow \text{OBJ})$

It would be desirable if this could be a general constraint on Romance verbs, but there is a well-known class of verbs that violate it. These are the verbs that take dative objects with no accompanying accusative 'direct' object, such as, in Catalan, *cridar* 'shout at':

(55) En Ferran li crida.
the Ferran him.DAT shouts
'Ferran shouts at him.' (Alsina 1996, p. 172)

I therefore propose that (54) is a specific constraint on causative verbs.

This analysis can also manage the 'long passives', that are found in Italian and Catalan, but not in Spanish or French (Alsina 1996, p. 187). According to Alsina (p.c.), passives of causatives don't sound truly natural, but sentences such as these below are possible:

(56) a. El pont ha estat fet enderrocar a un especialista
the bridge has been made repair.INF by a specialist
'Someone has had the bridge repaired by a specialist (the repairer).' (c.f. (2) of Alsina 1996, p. 187)

b. El poema ha estat fet llegir al nen.
the poem has been made read.INF a.the boy
'Someone has had the poem read by the boy.'

In these cases, if the causative verb is passivized, the ARG_2 of the lower verb can be realized as the SUBJ, in accordance with the usual mapping rules. There is more to be said about valence alternation in restructuring-style causatives, but this should be enough to establish that combining KFMT with the present theory about f-structure is a viable prospect.

We have now shown how long passives, fusion of argument arrays, and clitic climbing work in our account, these being the three main aspects of the monoclausality that is the problematic feature of these constructions. These are all consequences of the claim that they have a single array of grammatical relations, shared across all the levels of complex predicate constructions. We now briefly consider Lowe's 2015 analysis of Hindi, which shows some similar phenomena in its causative constructions.

4.5 *Lowe's 2015 analysis of Hindi*

As we mentioned earlier, Lowe thoroughly and cogently critiques all previous analyses of restructuring complex predicates, relieving us of this rather demanding task. He then presents his own treatment of Hindi, where the main and all the light verbs correspond to the same f-structure, but the meaning-constructors introduced by the light verbs apply to each other and to that of the main verb so as to build a hierarchical interpretation. This works well for Hindi, and is in fully standard LFG + glue, but has some problems. The first, which we have

already discussed, is that there are too many restructuring verbs in Catalan to plausibly treat them as not having PRED-features, but only being distinguished by some other kind of feature.

Another issue is that he says nothing about form-determination. As in Romance, different light verbs select different forms on their (in Urdu, linearly preceding) semantic complements. So completive *le* ('take') takes a (preceding) bare infinitive complement, while permissive *de* 'give' takes an oblique infinitive. This could be easily accommodated with the 'classic' m-projection from c-structure, but as we have noted, this proposal does not seem to find uses beyond the kinds of facts for which it was originally devised, and its subsequent adaptation to an m-structure that comes off f-structure is more complicated (I assume that having two kinds of m-structure, one from c-structure, the other from f-structure, should be rejected unless there is overwhelming evidence in favor of it). Furthermore, the worked out adaptation, for French (Frank and Zaenen 2004), seems to assume a flat sequence of V's, while Butt (1995) argues that Urdu also has both these and also VP complement clause union structures, like those of Spanish and Catalan, but with the order reversed.

The last and most serious problem is that, as Lowe discusses on his pp. 438–441, his analysis cannot account for the dependence of the semantic interpretation on the hierarchical structure, because it depends on composing meaning-constructors connected to f-structure, which on his analysis of these constructions is flat rather than hierarchical. He accepts this as a deficiency, and observes that the attempt in Andrews and Manning (1999) to overcome it involved major changes to LFG, and furthermore didn't address the problem of adjective scope addressed in Andrews and Manning (1993). He is therefore willing to leave it as a 'long term problem'. The proposal of this paper, however, does overcome both problems, and with only small modifications to the current LFG framework, depending on which recent independent proposals are regarded as already accepted.

5 CONCLUSION

I have proposed modest extension to pre-existing ideas in LFG to solve some longstanding problems with the capacity of the theory. In terms of the formal architecture, it might be that there is no actual change

at all, but only a change in the default structure-function mapping, with certain (possibly most or even all) kinds of c-structure heads marked by default with an $\downarrow \in \uparrow$ annotation rather than $\uparrow = \downarrow$. A remaining question is the treatment of PRED-features. For the analysis of restructuring predicates, we need PRED to be non-distributive, but this is not necessary for our analysis of modification, and Frank (2006) provides evidence from asymmetric coordination in German that PRED is distributive. If we decide the PRED is distributive, we can amend the analysis of Catalan by adding PRED to the undersharing specification of rule (39).

Observe that while the necessity for default nondistributivity of ADJUNCTS consists of subtle facts of interpretation and relatively rare grammatical phenomena, the stipulated nondistributivity of the verbal form features and possibly PRED is necessary to provide a reasonable analysis of the overt form of plentiful data, given the existence of clitic climbing and the other indications of 'monoclausality' (on this analysis, distribution/sharing of grammatical relations). So there would be a substantial Poverty of the Stimulus problem for stipulated nondistributivity of ADJUNCTS, but it is less serious for the stipulated nondistributivity of certain morphological features and maybe PRED, due to the more overt character of the evidence.

ACKNOWLEDGEMENTS

I would like to acknowledge Alex Alsina for a great deal of advice and judgements about Catalan, as well as an invitation to present the material at UPF Barcelona. Section 4 reproduces his analysis of Catalan data within a different framework. I am also indebted to Andrew Morrison, Mary Dalrymple, Adam Przepiórkowski, the members of the LFG glue semantics discussion group, and three anonymous JLM reviewers for useful feedback. I also acknowledge Christopher Manning, with whom I co-developed the precursors to the present proposal. Any errors, as usual, are mine.

APPENDICES

A THE REPRESENTATION OF DISTRIBUTED ATTRIBUTES

Although the concept of distributive attribute has been around for some time, there does not appear to have been any explicit attempt to work it into the LFG solution algorithm as presented originally in Kaplan and Bresnan (1982, 273–274). Suppose we are processing the functional description for example (4). At some point we will encounter the annotation saying that the f-structure of the subject NP *Mary* is the SUBJ of the f-structure of the whole sentence. At this point, we might or might not know that this f-structure is a hybrid object, and if we do know this, we might or might not know what all of its members are. In order to be independent of processing order, the algorithm needs to proceed smoothly and monotonically in all cases. I suggest that a way to achieve this is to represent the f-structure of the subject explicitly as the SUBJ-value of the entire clausal f-structure, i.e. at the top level of the set-inclusion structure, as in (4b). Then, when the information to the effect that some f-structure is a member of the f-structure of the whole sentence beomes available, the information about distributive attributes of the whole can be copied into it.

On the other hand, there is a different situation that can arise when the value of some distributive attribute such as TENSE is specified the same way internally in each member. A reasonable strategy would be to do nothing, unless a constraining specification wants to check the value of the attribute in the entire structure; in this event, one would then check its value in the members. I doubt that doing more than this would facilitate processing. This leads to a slight discrepancy in the representation of distributive attributes in different situations, although I don't see how that would create any real problems.

An anonymous referee points out that constraining specifications bring out a difference between the attribute-based account of distributivity from Dalrymple (2001) and the property-based one of Dalrymple and Kaplan (2000), which is that under the latter conception, an existential constraint such as (f TENSE) will be satisfied if every member of a hybrid object f has some TENSE value, even if they are not all the same, while under the former, it won't be. This is an interesting formal difference, but is unlikely to produce an empirically discernable

effect, since we can always propose that TENSE is a structured attribute where there is always at a minimum a common sub-attribute such as +. The implemention suggested in the previous paragraph whereby constraints are only checked without any sharing being effected might allow the two conceptions to be combined in practice.

Stipulating non-distributivity of a compound attribute such as INDEX NUM, while INDEX GEND is to remain distributive, requires more complex arrangements than simple ones, but is not impossible.

B GLUE SEMANTICS FOR ADJECTIVES

Here I will briefly show how to adapt Dalrymple's (2001) glue semantics for attributive adjectives to the present proposal. Sample constructors for the two modal adjectives *former* and *confessed* are:

(57) In both below, $\%G = (\text{ADJUNCTS} \in \uparrow)$:

$$\lambda P x.\mathit{Former}(P(x)) \ : \ [(\%G \in_\sigma \text{VAR}) \multimap (\%G_\sigma)] \multimap (\%G_\sigma \ \text{VAR}) \multimap \%G_\sigma$$

$$\lambda P x.\mathit{Confess}(x, P(x)) \ : \ [(\%G \in_\sigma \text{VAR}) \multimap (\%G \in)_\sigma] \multimap (\%G_\sigma \ \text{VAR}) \multimap \%G_\sigma$$

The changes from Dalrymple's (2001, p. 264) formulation are that the glue-side terms are a bit more complex in order to be able to apply the adjective meaning to that of the sister $\overline{\text{N}}$ and ascribe the result to the mother $\overline{\text{N}}$, and also the RESTR attribute is eliminated from the semantic projection, because it has no clear function. VAR should also be reconsidered, and its relationship to the widely proposed INDEX and CONCORD attributes established, but I won't do this here.

For intersectives, and similar, Dalrymple proposes two constructors, the first of which can be retained unaltered (other than the removal of RESTR), here illustrated by the one for *Swedish*:

(58) $\lambda x.\mathit{Swedish}(x) \ : \ (\uparrow_\sigma \text{VAR}) \multimap \uparrow_\sigma$

This is very close to what is needed for predicate adjectives. The other constructor that Dalrymple proposes is more complex, and effects the intersection of the adjectival meaning with the nominal meaning as constructed so far. Our version of it would be:

(59) $\%G = (\text{ADJUNCTS} \in \uparrow)$:
$\lambda PQx.P(x) \wedge Q(x)$:
$[(\uparrow_\sigma \text{VAR}) \multimap \uparrow_\sigma] \multimap [(\%G \in_\sigma \text{VAR}) \multimap (\%G \in)_\sigma] \multimap (\%G_\sigma \ \text{VAR}) \multimap \%G_\sigma$

Andrews (2010) suggests that this is a 'universal' meaning-constructor, similar in effect to the type-shifting rules widely employed in formal semantics.

REFERENCES

Artemis ALEXIADOU (2017), Gender and Nominal Ellipsis, in Nicholas LACARA, Keir MOULTON, and Anne-Michelle TESSIER, editors, *A Schrift to Fest Kyle Johnson*, pp. 11–22, Open Access Publications 1, University of Massachussetts, Amherst MAURL: `https://scholarworks.umass.edu/linguist_oapubs/1/`.

Alex ALSINA (1996), *The Role of Argument Structure in Grammar*, CSLI Publications, Stanford, CA.

Alex ALSINA (1997), A Theory of Complex Predicates: Evidence from Causatives in Bantu and Romance, in Alex ALSINA, Joan BRESNAN, and Peter SELLS, editors, *Complex Predicates*, pp. 203–246, CSLI Publications, Stanford, CA.

Elena ANAGNOSTOPOULOU (2003), *The Syntax of Ditransitives*, de Gruyter, Berlin.

Avery D. ANDREWS (1983), A Note on the Constituent Structure of Modifiers, *Linguistic Inquiry*, 14:695–697.

Avery D ANDREWS (2007), Glue Semantics for Clause-Union Complex Predicates, in Miriam BUTT and Tracy Holloway KING, editors, *The Proceedings of the LFG '07 Conference*, pp. 44–65, CSLI Publications, Stanford CA.

Avery D. ANDREWS (2010), 'Grammatical' vs. 'Lexical' Meaning Constructors for Glue Semantics, in Yvonee TREIS and Rik De BUSSER, editors, *Selected Papers from the 2009 Conference of the Australian Linguistic Society*, The Australian Linguistic Society, URL: `http://www.als.asn.au/proceedings/als2009/andrews.pdf`.

Avery D. ANDREWS and Christopher D. MANNING (1993), Information-Spreading and Levels of Representation in LFG, Technical Report CSLI-93-176, Stanford University, Stanford CA, `http://nlp.stanford.edu/~manning/papers/proj.ps`; viewed 18 Feb 2018.

Avery D. ANDREWS and Christopher D. MANNING (1999), *Complex Predicates and Information Spreading in LFG*, CSLI Publications, Stanford, CA.

Ash ASUDEH (2005), Control and Resource Sensitivity, *Journal of Linguistics*, 41:465–511.

Ash ASUDEH and Richard CROUCH (2002), Coordination and Parallelism in Glue Semantics: Integrating Discourse Cohesion and the Element Constraint, in *Proceedings of the LFG02 Conference*, in Miriam BUTT and Tracy Holloway KING, editors, pp. 19–39, CSLI Publications, Stanford, CA.

Ash ASUDEH, Gianluca GIORGOLO, and Ida TOIVONEN (2014), Meaning and Valency, in *Proceedings of the LFG14 Conference*, in Miriam BUTT and Tracy Holloway KING, editors, pp. 68–88, CSLI Publications.

Oleg BELAYEV (2013), Optimal Agreement in Dargwa: Person at m-structure, in Tracy Holloway KING and Miriam BUTT, editors, *Proceedings of the LFG13 Conference*, pp. 90–110, CSLI Publications, Stanford, CA.

Oleg BELAYEV, Mary DALRYMPLE, and John J. LOWE (2015), Number Mismatches in Coordination: An LFG Analysis, in Tracy Holloway KING and Miriam BUTT, editors, *Proceedings of the LFG15 Conference*, pp. 26–46, CSLI Publications, Stanford, CA.

Joan Wanda BRESNAN, Ronald M. KAPLAN, and Peter PETERSON (1985), Coordination and the Flow of Information through Phrase Structure, Unpublished manuscript.

Miriam BUTT (1995), *The Structure of Complex Predicates in Urdu*, CSLI Publications, Stanford CA, originally Stanford Ph.D. dissertation, 1993.

Miriam BUTT, Tracy Holloway KING, Mará-Eugenia NIÑO, and Frédérique SEGOND (1999), *A Grammar-Writer's Cookbook*, CSLI Publications, Stanford CA.

Miriam BUTT, Maria Eugenia NIÑO, and Frédérique SEGOND (1996), Multilingual Processing of Auxiliaries within LFG, in Dafydd GIBBON, editor, *Natural Language Processing and Speech Technology*, pp. 111–122, Mouton de Gruyter, Berlin.

Greville CORBETT (1991), *Gender*, Cambridge University Press, Cambridge.

Greville G. CORBETT (1979), The Agreement Hierarchy, *Journal of Linguistics*, 15:203–224.

Mary DALRYMPLE, editor (1999), *Syntax and Semantics in Lexical Functional Grammar: The Resource-Logic Approach*, MIT Press, Cambridge MA.

Mary DALRYMPLE (2001), *Lexical Functional Grammar*, Academic Press, Cambridge MA.

Mary DALRYMPLE and Ronald M KAPLAN (2000), Feature Indeterminacy and Feature Resolution, *Language*, 76:759–798.

Jamie FINDLAY (2016), Mapping Theory without Argument Structure, *Journal of Language Modelling*, 4:293–338.

Annette FRANK (2006), A (Discourse-) Functional Analysis of Asymmetric Coordination, in *Intelligent Linguistic Architectures: Variations on a Theme by Ronald M. Kaplan*, in Miriam BUTT, Mary DALRYMPLE and Tracy Holloway KING, editors, pp. 259–285, CSLI Publications, Stanford CA.

Annette FRANK and Annie ZAENEN (2004), Tense in LFG: Syntax and Morphology, in Louisa SADLER and Andrew SPENCER, editors, *Projecting Morphology*, pp. 23–66, CSLI Publications, Stanford CA.

Ray S. JACKENDOFF (1977), $\overline{X}$*-syntax*, MIT Press, Cambridge MA.

Ronald M. KAPLAN and Joan BRESNAN (1982), Lexical-Functional Grammar: a Formal System for Grammatical Representation, in Joan BRESNAN, editor, *The Mental Representation of Grammatical Relations*, pp. 173-–281, MIT Press, Cambridge MA, also in Dalrymple *et al.*, editors, 1995 *Formal Issues in Lexical-Functional Grammar*, CSLI Publications, pp. 29–130; page number references to 1982 version.

Ronald M. KAPLAN and John T. MAXWELL (1988), Constituent Coordination in Lexical-Functional Grammar, in *Proceedings of COLING-88, vol I*, pp. 303–305, also in Dalrymple, Kaplan, Maxwell and Zaenen (1995), pp 199–210.

Ronald M. KAPLAN and Jürgen WEDEKIND (1993), Restriction and Correspondence-Based Translation, in *Proceedings of the Sixth European Converence of the Association for Computational Linguistics*, pp. 193–202, Utrecht, URL: `http://www.aclweb.org/anthology/E93-1024`.

Anna KIBORT (2013), Objects and Lexical Mapping Theory [Abstract], in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG13 Conference*, CSLI Publications, Stanford CA.

Tracy Holloway KING and Mary DALRYMPLE (2004), Determiner Agreement and Noun Conjunction, *Journal of Linguistics*, 40:69–104.

Jonas KUHN (2001), Resource Sensitivity in the Syntax-Semantics Interface and the German Split NP Construction, in W. Detmar MEURERS and Tibor KISS, editors, *Constraint-Based Approaches to Germanic Syntax*, CSLI Publications, Stanford CA.

Idan LANDAU (2016), DP-internal Semantic Agreement: A Configurational Analysis, *Natural Language and Linguistic Theory*, pp. 975–1020.

John LOWE (2015), Complex Predicates: an LFG + glue Analysis, *Journal of Language Modelling*, 3:413–462.

Christopher D. MANNING (1996), Romance Complex Predicates: In Defence of the Right-Branching Structure, paper presented at the Workshop on Surfaced-Base Syntax and Romance Languages, 1996 European Summer School on Logic, Language and Information, Prague. Draft available at URL: `https://nlp.stanford.edu/~manning/papers/right-paper.pdf`; viewed 18 Feb 2018.

Jason MERCHANT (2014), Gender Mismatches under Nominal Ellipsis, *Lingua*, 151:9–32.

Rachel NORDLINGER and Louisa SADLER (2008), From Juxtaposition to Incorporation: an Approach to Generic-Specific Constructions, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG08 Conference*, CSLI Publications, Stanford CA.

Sarah OUWAYDA (2014), *Where Number Lies: Plural marking, numerals, and the collective–distributive distinction*, Ph.D. thesis, USC, Los Angeles.

Barbara PARTEE (2010), Privative Adjectives: Subsective plus Coercion, in Rainer BÄUERLE, Uwe REYLE, and Thomas Ede ZIMMERMANN, editors, *Presuppositions and Discourse: Essays offered to Hans Kamp*, pp. 273–285, Emerald Group Publishing, Bingley UK.

David PESETSKY (2013), *Russian Case Morphology and the Syntactic Categories*, MIT Press, Cambridge MA.

Zorica PUŠKAR (2017), *Hybrid Agreement Modelling Variation, Hierarchy Effects and ϕ-feature Mismatches*, Ph.D. thesis, University of Leipzig, Leipzig, `http://ling.auf.net/lingbuzz/003795`.

Louisa SADLER and Douglas J. ARNOLD (1994), Prenominal Adjectives and the Phrasal/Lexical Distinction, *Journal of Linguistics*, 30:187–226.

Louisa SADLER and Rachel NORDLINGER (2010), Nominal Juxtaposition in Australian Languages: an LFG Analysis, *Journal of Linguistics*, 46:415–452.

Jaume SOLÀ (2002), Clitic Climbing and Null Subject Languages, *Catalan Journal of Linguistics*, 1:225–255.

Nikolaos VELEGRAKIS (2011), *The Syntax of Greek Polydefinites*, Ph.D. thesis, University College London, London, `http://discovery.ucl.ac.uk/1302548/1/1302548.pdf`; viewed Feb 18, 2018.

Jean-Roger VERGNAUD (1974), *French Relative Clauses*, Ph.D. thesis, MIT, Cambridge MA.

Stephen WECHSLER (2011), Mixed Agreement, the Person Feature, and the Index/Concord distinction, *Natural Language and Linguistic Theory*, pp. 999–1031.

Stephen WECHSLER and Larisa ZLATIĆ (2000), A Theory of Agreement and its Application to Serbo-Croatian, *Language*, 76:799–832.

Stephen WECHSLER and Larisa ZLATIĆ (2003), *The Many Faces of Agreement*, CSLI Publications, Stanford CA.

Temporal predictive regression models for linguistic style analysis

Carmen Klaussner and Carl Vogel
School of Computer Science and Statistics,
Trinity College Dublin

ABSTRACT

This study focuses on modelling general and individual language change over several decades. A timeline prediction task was used to identify interesting temporal features. Our previous work achieved high accuracy in predicting publication year, using lexical features marked for syntactic context. In this study, we use four feature types (character, word stem, part-of-speech, and word n-grams) to predict publication year, and then use associated models to determine constant and changing features in individual and general language use. We do this for two corpora, one containing texts by two different authors, published over a fifty-year period, and a reference corpus containing a variety of text types, representing general language style over time, for the same temporal span as the two authors. Our linear regression models achieve good accuracy with the two-author data set, and very good results with the reference corpus, bringing to light interesting features of language change.

Keywords:
language change,
style analysis,
regression

1 INTRODUCTION

Statistical style analysis or 'stylometry' is the automatic analysis of authorial style, usually investigating the frequency of occurrence of specific features in a given author's works. Features with consistent frequencies are assumed to be representative of that author, and features are also considered discriminative if other comparable authors use them with consistently different frequencies. This type of analy-

sis is known as synchronic analysis, as it disregards composition or publication dates.

However, this is a simplification, since most writers compose over time spans of 20–40 years, where they not only undergo individual stylistic development, but also bear witness to general contemporaneous language change. These two types of temporal influences can cause synchronic analyses to be misinterpreted. Thus, as already discussed by Daelemans (2013), unless style is found to be invariant for an author and does not change with age and experience, temporality can be a confounding factor in stylometry and authorship attribution. For this reason, diachrony presents an important aspect of style analysis, not only to disambiguate synchronic analyses of style, but also in its own right by modelling language change over time.

In this work, we examine language change in two literary authors, as well as the corresponding background language change during the same time period. Specifically, we are interested in features that are attested in each time slice of the diachronic corpus studied. We refer to this subset of features that appear in all samples as 'constant' features. This classification captures occurrence patterns rather than variation in terms of relative frequencies, which may or may not change over the time intervals examined. In order to identify salient constant features that exhibit change over time, we refer to a temporal prediction task based on the features' relative frequencies.

This extends our previous work on predicting the publication year of a text using syntactic word features (Klaussner and Vogel 2015).[1] That study considered a data set comprising works by two authors from the 19th to the 20th century, as well as a data set based on a reference corpus, and sampled features that appeared in many, but not necessarily all, time slices. For the two-author data set, a root-mean-square error (RMSE) of 7.2 years[2] on unseen data (baseline: 13.2) was

[1] These are lexical features that have been marked for syntactic function to differentiate between lexical representations that can appear in different syntactic contexts (see Section 4.2).

[2] Hereafter, when we report RMSE, we take the units to be years and do not repeat the unit. This is to be understood with the caveat that the data are processed using only integer values of years. Temporal prediction for any text cannot be wrong by '7.2 years', but rather by seven or eight years. The RMSE is an aggregate.

obtained, whereas the model built on the larger reference data set obtained an RMSE of 4 on unseen data (baseline: 17). While the current work is similar in that it uses the same data sets and the same general prediction task, it is different in that achieving 'high accuracy' of prediction is not the main objective here. Although we report our results and compare them to those from the earlier study, the prediction task is primarily used as a means to determine what is stable and what changes in individual and general language use over time.[3] Hence, the purpose is not the pursuit of a perfect temporal classifier, but rather to understand 'typical' distributions of linguistic feature categories during an author's lifetime. This change must also be understood in relation to the effects of ageing on language production, as explored for instance by Pennebaker and Stone (2003). Features that are not constant in the sense analysed here are also important. We focus on constant features, because if they are used in each time slice throughout an author's career, then they are probably integral to that author's style, making the relative frequencies of such features across time slices interesting to explore.

The contribution of this new study is the analysis of language change using an extended feature set, adding character,[4] word stem, and syntactic (part-of-speech tag) features to the previous set, which consisted only of syntactic word features. In addition, rather than considering only unigram size, this study analyses all n-gram sizes up to length four. Therefore, one of the questions investigated as part of this work is whether (and to what extent) the more linguistically informative features, such as syntactic word n-grams, exhibit more dramatic change than lexicographic and part-of-speech features. We present our own method for reasoning about temporal change in constant linguistic features, using standard techniques from regression analysis, particularly parameter shrinkage.[5] We find that the best predictive values common to the works by the two authors and the reference corpus are word stem, and POS bigrams and trigrams, which also account for

[3] The data sets for the two authors are analysed both separately and together.

[4] This feature type covers alphanumeric characters, punctuation, and spaces.

[5] The resulting set of features identified is a specific subset of features that are both constant and have a linear relationship with the response variable over time, i.e. a change in trend rather than in periodicity. Non-linear patterns or estimation may also be interesting, but our focus is different here.

most shared model predictors. In terms of language change, with the help of our regression models, we identified several differences between the reference corpus and the works by the two authors.

The remainder of this article is structured as follows: Section 2 outlines previous work in the area; Section 3 discusses methods; Section 4 presents the data sets, preprocessing steps, and feature types; Section 5 discusses the general experimental setup and the experiments themselves. Section 6 reports and analyses the salient features of the models. Section 7 discusses the results, and Section 8 concludes this work.

2 RELATED WORK

Studies in the field of style analysis or 'stylometry' focus on different sub-tasks, such as authorship attribution; i.e. given an unknown document and several candidate authors, the task is to decide which candidate is most likely to have authored the document. This problem can be studied in a closed-class or open-class scenario. The former assumes that the true author is among the set of candidates, rendering the task of determining who authored the document in question simpler than in the open-class variant, where the set of candidates may or may not contain the true author. Open-class authorship attribution has been studied for instance by Koppel *et al.* (2011), who consider authorship attribution in the presence of what they conceive are the three most common deterrents to using common authorship techniques, i.e. possibly thousands of known candidate authors, the author of the anonymous text not being among the candidates, and the 'known-text' for each candidate and/or the anonymous text being very limited. Considering a set of blog posts (extracting 2,000 words for the known text and a 500-word-long test snippet), they use a similarity-based approach (cosine similarity) on space-free character tetragrams. The task is to find the author of a given text snippet, based on evidence from varying feature sets, the rationale being that only the right author is going to be consistently similar to his or her own 'unknown' piece. An author is selected only if above a particular proportion or threshold, otherwise the method returns a 'Don't know' answer. Unsurprisingly, a greater number of feature sets and a closed-candidate set yield greater accuracy, i.e. 87.9% precision with 28.2% recall. In

the closed-candidate setting, reducing the number of candidates improves accuracy (e.g. 1,000 candidates yields 93.2% precision with 39.3% recall), whereas in the open-class setting, having fewer candidates actually introduces problems, in that an author might end up being chosen erroneously, because there is less competition. Overall, Koppel *et al.* (2011) find that their methods achieve passable results even for snippets as short as 100 words, but note that there is still no satisfactory solution for the case of a small open-candidate set and limited anonymous text.

Another general variant of the attribution problem is commonly referred to as 'Authorship verification', which requires determining whether a piece of text has been written by a specific author. This has been considered by Koppel *et al.* (2007), for instance, who show that the task of deciding whether an author has written a particular text can be accurately determined by iteratively removing the set of best features from the learning process: the differences between two texts by the same author are usually only reflected in a relatively small number of features, causing accuracy to drop much faster and more dramatically than when the texts were not written by the same person. In contrast, 'Author profiling', which involves predicting an author's characteristics, such as gender, age or personality traits, based on a particular text, has been studied extensively as part of the PAN competitions (e.g. see Rosso *et al.* 2016). While the predicted variable varies by task, what is common to the studies above as well as to our own is the use of relative frequencies of some feature to predict the variable of interest, using similarity-based or statistical methods.

However, while the general scenario is the same, diachronic studies differ in that they take into account the temporal ordering of an author's works, seeking to reveal temporal changes within his or her style rather than changes between authors or between different texts by the same author. A few works focus more specifically on temporality in style analyses. Previous work by Smith and Kelly (2002) investigates the question of whether vocabulary richness remains constant over time, by examining measures of lexical richness across the diachronic corpora of three playwrights (Euripides, Aristophanes, and Terence). The plays are divided into standardized non-overlapping blocks, each being analysed for certain properties pertaining to lexical richness, such as vocabulary richness, pro-

portion of *hapax legomena*, and repetition of frequently appearing vocabulary. In addition to testing the constancy of these properties over time, weighted linear regression is used to test associations between these measures and the time of a play's first performance. For this, the property's value in a particular text block is used as response, and time of performance is used as predictor.[6] Results show that Aristophanes' use of *hapax legomena* appears to have decreased over time. Interestingly, one of his earlier works, *Clouds*, which was subjected to redrafting after the first staging, but for which the finishing date is unknown, is predicted to originate towards the end of the playwright's life, indicating that revisions might have been made at a much later stage. Our work here also uses linear regression, but rather than using time as predictor, we investigate to what extent pooled information from several features can accurately predict a text's publication year. The study presented by Hoover (2007) considers language change in Henry James' style with respect to the 100–4,000 most frequent word unigrams, using methods such as 'Cluster Analysis', 'Burrows' Delta', 'Principal Component Analysis', and 'Distinctiveness Ratio'.[7] Three different divisions, into early (1877–1881), intermediate (1886–1890), and late style (1897–1917), emerge from the analysis.[8] However, rather than being strict divisions, there seem to be gradual transitions, with the first novels of the late period being somewhat different from the others, suggesting that it might be interesting to conduct a continuous analysis of style in James' works. Thus, in contrast to the previous study, the work we present here focuses on a more graduated interpretation of style over time, with yearly intervals rather than classification into

[6] In order to perform inverse prediction, i.e. predicting the date of an unknown work by the measure, the authors draw a horizontal line at y, with y corresponding to the measure's average in the text and look at the intersection with the estimated regression line.

[7] Distinctiveness Ratio: Measure of variability defined by the rate of occurrence of a word in a text divided by its rate of occurrence in another. Principal Component Analysis (PCA) is an unsupervised statistical technique to convert a set of possibly related variables to a new uncorrelated representation, i.e., principal components.

[8] The same divisions have also been identified by literary scholars (Beach 1918).

different periods along the timeline of the author's works. Our work on temporal prediction (Klaussner and Vogel 2015) considered the task of accurately predicting the publication year of a text through the relative frequencies of syntactic word features.[9] We used multiple linear regression models to predict the year when a text was published, for three data sets, the first containing texts by Mark Twain and Henry James, the second a mid 19th to early 20th century reference corpus, and a third one combining all data from the previous two sets. Although the data for the two authors had been kept separate to allow for potentially different levels between them, the models disregarding authorial source tended to be more accurate (RMSE of 7.2 vs. 8.0). While the reference corpus model performed well on its own test set (RMSE of 4), using it to predict publication year for the two authors was rather inaccurate (RMSE: 15.4 for Twain, and 20.3 for James). This suggests that the style of the two authors was rather different from general language, Twain's being somewhat more similar to it than James'. Combining all data leads to more accurate results (RMSE: 1.8), and model features and estimates suggest a marked influence of Twain and James on the model, in spite of their smaller data sets (for more detailed, quantitative results, see Section 5.3).

On the topic of suitable stylistic feature types in this context, Stamatatos (2012) compares the performances of the most frequent function words and character trigrams for the authorship attribution task. It is shown that character trigrams outperform word features, especially when training and test corpus differ in genre – they are also found to be more robust and effective when considering different feature input sizes. For this reason, we include character n-grams as a feature type here as well. In contrast to part-of-speech tags or word stems, character n-grams present a less linguistically motivated feature type, as writers would not be able to control the number of times a particular character is used to the same extent as they would be able to control their choice of particular syntactic constructions. Yet this feature type becomes more likely to bear meaning, as character n-gram size increases, approaching average word length.

[9] Syntactic word features are words marked for their syntactic context. This is explained in more detail in Section 4.2.

3 METHODS

This section discusses the methods used in this work, beginning with temporal regression models (Section 3.1), and continuing with evaluation techniques for these predictive models (Section 3.2).

3.1 *Temporal regression models*

The analysis of data over time probably has its most prominent usage in quantitative forecasting analysis, which involves the (quantitative) analysis of how a particular variable (or variables) may change over time and how that information can be used to predict its (or their) future behaviour, thus inherently assuming that some aspects of the past continue in the future, known as the 'continuity assumption' (Makridakis *et al.* 2008). Thus, a future value of a variable y is predicted using a function over some other variable values. These other variable values could be composed in two different ways, pertaining either to the use of a 'time-series' model or an 'explanatory' model. When considering a time-series model, the assumption is that one can predict the future value of the variable y by looking at the values it took at previous points in time and the possible patterns this would show over time. In contrast, for prediction, explanatory models focus less on interpreting previous values of the same variable, and more on the relationship with other variables at the same point in time. Consequently, the prediction of a variable y, using explanatory models, is based on a function over a set of distinct variables: $x_1, x_2, \ldots, x_{p-1}, x_p = X$, with $y \notin X$, at the same time point $t : \{t \in 1, \ldots, n\}$, and some error term: $y_t = f(x_{1t}, \ldots, x_{2t}, \ldots, x_{p-1t}, \ldots, x_{pt}, error)$.

The general model for this is shown in Equation (1), predicting variable y, where $\hat{y}_t$ refers to the estimate of that variable at a particular time instance $t : \{t \in 1, \ldots, n\}$, β_0 refers to the intercept, and β_p to the pth coefficient of the pth predictor x_{pt}.

$$\hat{y}_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_p x_{pt} \tag{1}$$

In the present case, the year of publication is always set as the response variable, so that a model based on syntactic unigrams (relative frequencies) for the year 1880 could be defined in the following way: $\hat{y}_{1880} = \beta_0 + \beta_1(NN_{1880}) + \beta_2(NP_{1880}) + \beta_3(IN_{1880})$.

Regression models are customarily evaluated using the residual sum of squares (RSS): given predicted values $\hat{y}_i$ computed by the

model and observed values y_i, the RSS measures the difference between them. The smaller the RSS, the greater the amount of variation of y values around their mean that is explained by the model. This is known as the 'ordinary least squares' (OLS) fit, a model selection criterion that also forms the basis of evaluation measures, such as the root-mean-square error (RMSE) (see Section 3.2).

In this work, rather than applying models based only on least squares regression, we employ so-called 'shrinkage' models that offer an extension to regular OLS models by additionally penalizing coefficient magnitudes, thus aiming to keep the model from overfitting the data. Specifically, we use the 'elastic net', which is a combination of the two most common types of shrinkage, 'lasso' and 'ridge' regression (Zou and Hastie 2005). The elastic net penalizes both the L_1 and L_2 norms,[10] causing some coefficients to be shrunk (ridge) and some to be set to zero (lasso), with the exact weighting between the two also being subject to tuning. In addition, the elastic net tends to select groups of correlated predictors rather than discarding all but one from a group of related predictors, as is common when using only the lasso technique. The entire cost function is shown in Equation (2). As with the lasso and ridge regression, $\lambda \geq 0$ controls finding a compromise between fitting the data and keeping coefficient values as small as possible, while the elastic net parameter α determines the mix of the two penalties, i.e. how many features are merely shrunk as opposed to being completely removed.

$$\max_{\{\beta_{0,k},\beta_k \in \mathbf{R}^p\}_1^K} \left[\sum_{i=1}^{N} \log \Pr(g_i|x_i) - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} (\alpha|\beta_{kj}| + (1-\alpha)\beta_{kj}^2) \right] \tag{2}$$

There are numerous advantages to using shrinkage models, and the elastic net estimation in particular, such as built-in feature selection and more robust and reliable coefficient estimation. This is discussed in more detail for instance by James *et al.* (2013, pp. 203–204) and Friedman *et al.* (2001, pp. 662–663).

3.2 *Evaluation*

The 'root-mean-square error' (RMSE) is one of the measures that can be used for the purpose of evaluating linear regression models: it is

[10] $\|\beta\|_1$: $\sum_i |\beta_i|$ and $\|\beta\|_2^2$: $\sum_i \beta_i^2$

defined as the square root of the variance of the residuals between outcome and predicted value and thus provides the standard deviation around the predicted value, as shown in Equation 3.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \tag{3}$$

The advantage over the more general 'mean-square error' (MSE) is that RMSE computes deviations in predictions on the same scale as the data. However, due to the squaring, assigning more weight to larger errors, the RMSE is more sensitive to outliers.

4 DATA

The following section presents the data sets (Section 4.1), followed by feature types (Section 4.2), and finally, data preparation (Section 4.3).

4.1 *Data sets*

The data for this study originates from three separate sources: works by two American authors, Mark Twain and Henry James, and a reference corpus for American English, from 1860 to 1919.

Mark Twain and Henry James were chosen for this analysis because both were prolific authors writing over a similar time span, from the late 19th to the early 20th century. The study presented by Hoover (2007), mentioned in Section 2, provided the first evidence that a temporal analysis of James' work might be fruitful; other sources (Beach 1918; Canby 1951) indicated that it might be interesting to study works by Henry James and Mark Twain, two highly articulate and creative writers, contrasting in temperament and in their art (Canby 1951, p. xii), yet each conscious of the other (Brooks 1920; Ayres 2010). Considering individual authors might be more interesting from an interpretative viewpoint, in that the phenomena observed are more likely to be directly attributable to the author(s) examined. However, one needs a reference corpus representing 'average' style to know what importance to assign to a particular phenomenon. For instance, one might discover a decrease in usage of a particular feature over time for Twain and James; if the same feature also decreased in usage in general, this discovery would not necessarily be noteworthy. While

both individual and general language change are of interest in their own right, they also provide comparative information about the relative importance of the features observed, indicating whether particular events are likely to be unusual.

For each of the two authors, we compiled a separate data set of their main works.[11] Table 1 shows the data for Henry James, and Table 2 that for Mark Twain. The texts were collected from the *Project Gutenberg*[12] and the *Internet Archive*[13] selecting the earliest editions available. The reference corpus was assembled by taking an extract from *The Corpus of Historical American English* (COHA; Davies 2012).[14] The COHA is a 400-million-word corpus, containing samples of American English from 1810–2009, balanced in size, genre and sub-genre in each decade (1,000–2,500 files each). It contains balanced language samples from fiction, popular magazines, newspapers and non-fiction books, which are again balanced across sub-genres, such as drama and poetry.[15] The COHA data were compiled from different sources, some of which were already available as part of existing text archives (e.g., *Project Gutenberg* and *Making of America*), whereas others were converted from PDF images, or scanned from printed sources. The corpus allows analysis of linguistic change at different levels, i.e. lexical, morphological, syntactic, and semantic.

4.2 *Feature types*

For the experiments described in Section 5, we consider four different types of features, as well as various sequence sizes of these. Table 3 lists all feature types, ordered by increasing degree of specificity, with an example for unigrams, and one for trigrams.

The most general type is character n-grams, including punctuation and single spaces.[16] While the character n-grams reduce words

[11] In this case, 'main' is with reference to the size of the work in kilobytes, rather than in terms of literary importance. We use kilobytes instead of word count, as this gives a more precise indication of file size.

[12] `http://www.gutenberg.org/` – last verified March 2018.

[13] `https://archive.org/` – last verified March 2018.

[14] Free version available from: `http://corpus.byu.edu/coha/` – last verified March 2018.

[15] An Excel file with a detailed list of sources is available from: `http://corpus.byu.edu/coha/` – last verified March 2018.

[16] Multiple spaces were reduced to single spaces.

Table 1: Collected works for Henry James. Showing 'Title', the original publication date ('1st Pub.'), version collected ('Version'), 'Size' in kilobytes and 'Genre' type. The dashed lines indicate the boundaries for compression, i.e. which of the works are combined into one temporal interval (see Section 4.3 for discussion of the compression technique used)

Title	1st Pub.	Version	Size	Genre
The American	1877	1877	721	novel
Watch and Ward	1871	1878	345	novel
Daisy Miller	1879	1879	119	novella
The Europeans	1878	1879	346	novel
Hawthorne	1879	1879	314	biography
Confidence	1879	1880	429	novel
Washington Square	1880	1881	360	novel
Portrait of a Lady	1881	1882	1200	novel
Roderick Hudson	1875	1883	750	novel
The Bostonians	1886	1886	906	novel
Princess Casamassima	1886	1886	1100	novel
The Reverberator	1888	1888	297	novel
The Aspern Papers	1888	1888	202	novella
The Tragic Muse	1890	1890	1100	novel
Picture and Text	1893	1893	182	essays
The Other House	1896	1896	406	novel
What Maisie Knew	1897	1897	540	novel
The Spoils of Poynton	1897	1897	376	novel
In the Cage	1893	1898	191	novella
Turn of the Screw	1898	1898	223	novella
The Awkward Age	1899	1899	749	novel
Little Tour in France	1884	1900	418	travel writings
The Sacred Fount	1901	1901	407	novel
The Wings of the Dove	1902	1902	1003.7	novel
The Golden Bowl	1904	1904	1100	novel
Views and Reviews	1908	1908	279	literary criticism
Italian Hours	1909	1909	711	travel essays
The Ambassadors	1903	1909	890	novel
The Outcry	1911	1911	304	novel
*The Ivory Tower**	1917	1917	488	novel
*The Sense of the Past**	1917	1917	491	novel

'*' indicates unfinished works.

Table 2: Collected works for Mark Twain. Showing 'Title', the original publication date ('1st Pub.'), version collected ('Version'), 'Size' in kilobytes and 'Genre' type. The dashed lines indicate the boundaries for compression, i.e. which of the works are combined into one temporal interval (see Section 4.3 for discussion of the compression technique used)

Title	1st Pub.	Version	Size	Genre
Innocents Abroad	1869	1869	1100	travel novel
The Gilded Age: A Tale of Today	1873	1873	866	novel
The Adventures of Tom Sawyer	1876	1884	378	novel
A Tramp Abroad	1880	1880	849	travel literature
Roughing It	1880	1880	923	semi-autobiog.
The Prince and the Pauper	1881	1882	394	novel
Life on the Mississippi	1883	1883	777	memoir
The Adventures of Huckleberry Finn	1884	1885	586	novel
A Connecticut Yankee in King Arthur's Court	1889	1889	628	novel
The American Claimant	1892	1892	354	novel
The Tragedy of Pudd'nhead Wilson	1894	1894	286	novel
Tom Sawyer Detective	1896	1896	116	novel
Personal Recollections of Joan Arc	1896	1896	796	historical novel
Following the Equator	1897	1897	1000	travel novel
Those Extraordinary Twins	1894	1899	120	short story
A Double Barrelled Detective Story	1902	1902	103	short story
Christian Science	1907	1907	338	essays
Chapters from My Autobiography	1907	1907	593	autobiog.
*The Mysterious Stranger**	1908	1897–1908	192	novel

'*' indicates unfinished works.

and sentences to their orthography, the part-of-speech (POS) type generalizes them as sequences of syntactic types. Word stems present a more specific generalization of the simple word feature, but rather than capturing syntactic aspects, this type captures what lexical type of word (or sequence) was used, such as ⟨allud to⟩ in place of 'allude

Table 3: Feature types

n-gram type	Example	
	unigram	*trigram*
character	⟨c⟩	⟨ca,⟩
part-of-speech (POS)	⟨NP⟩	⟨IN DET NP⟩
word stem	⟨allud⟩	⟨to allud to⟩
syntactic word (lexical)	⟨like.IN⟩	⟨like.VB the.DET others.NNS⟩

to' or 'alludes to'.[17] The most specific is termed 'syntactic word' sequences, meaning words that have been marked for syntactic class, as in the case of 'like', which may be used as a preposition or a verb, depending on context. Compare *I'm like my father.* and *I like my father.*: in the first instance 'like' is used as a preposition, in the second it is used as a verb. Hence, for this feature type, each word is given the correct part-of-speech tag, thus allowing distinct features to be identified for words with more than one syntactic context, such as ⟨like.VB⟩ for verbal usage and ⟨like.IN⟩ for prepositional usage.

4.3 *Data preparation*

Before features could be extracted from the two authors' texts, each file had to be checked manually, to remove parts that were written at a different time from the main work, or introductions or comments not by the author, such as notes or introductions by editors. Following this, all source files were then searched (both automatically and manually) to remove unwanted formatting sequences and to normalize spacing.[18]

To extract both POS and syntactic word features, we used the TreeTagger POS tagger (Michalke 2014; Schmid 1994). The original word plus its tag is retained for syntactic word features, while for POS features, the original word is replaced by the POS tag.[19] After ex-

[17] The feature remains orthographic inasmuch as the stem differs from the lemma.

[18] The package *stylo* (Eder *et al.* 2013) was used to convert words into character sequences, while the *RTextTools* package (Jurka *et al.* 2012) was used to extract word stems.

[19] Punctuation and sentence endings are also included as features and in relativization. The POS tags assigned by the tagger to the individual word entity in its context are used to augment or replace the word entity. Individual entities within ⟨...⟩ are separated by a space.

traction, all feature types were then transformed to lowercase, as for this work we do not analyse features with respect to sentence boundaries. Finally, document-feature matrices were constructed for each type and n-gram size and relativized in the following way: for all of the analyses reported here, we compute relative frequencies to take into account any differences in the amount of text available for each year.[20] If more than one work was available for a particular year and authorial source, they were joined together and relativized as one text. For both the reference set and the two-author set, an ordinal variable 'year' was added for each experiment to mark the publication year of a text. The data sets for the two authors were joined into one set after relativization, with an additional categorical variable 'author' to mark which author composed the text. In some instances, both authors published work during the same year; the 'author' variable served to keep such cases separate. Thus, detecting differences in levels of relative frequency by author remains possible within the joint data set. Combined relativization might distort individual interpretation or create a shift towards the author with more data in a given year. The model is trained on 'combined' data, in the sense that there may be two relative frequencies contributing observations to one predictor variable. The categorical author variable may be added to the model, if the level for that predictor differs between James and Twain.

5 EXPERIMENTS

Section 5.1 addresses general experimental design, and model and parameter selection. The four feature types described in Table 3 are considered separately for the two data sets hereafter, with Section 5.2 presenting the results, and Section 5.3 comparing them with the previous study.

5.1 *Model computations*

Before the experiments, the same procedure was performed for all of the previously constructed document-feature matrices, to construct

[20] Long and rarer n-gram sequences could cause the data to become rather sparse and feature values could thus become computationally expensive. To overcome this challenge, memory-intensive processing steps were separated and simplified, using the R packages *bigmemory* (Kane *et al.* 2013) and *foreach* (Revolution Analytics and Weston 2014).

the input for each of the 32 models shown in Table 5. The data were first divided into training and test data using a 75/25 stratified split on the ordinal variable 'year' that we added at the previous step.[21] After that step, we extracted all constant features from the training set, i.e. the features appearing in all training set instances, which were then passed to the elastic net models.[22]

The final model was then computed by performing 10-fold cross-validation on the training data to find the 'best' α and λ parameters, deciding to what extent features were either shrunk or removed from the model as part of the elastic net configuration.[23] We defined the 'best' α and λ parameter estimates for a model as their combined global optimum. This optimum was then defined as the most parsimonious model within 1 standard error (SE) of the model with the lowest error, as defined by the MSE. By not choosing the best performing model, we could circumvent models that might be needlessly complex and thus somewhat balance prediction accuracy and model complexity. The evaluation parameter, RMSE, for the training and internal test set was computed by taking the model MSE and computing its square root. For evaluation on external data, we had to rebuild the training model manually from the model's coefficients.[24] Occasionally, the sets of constant features differed across training and (external) test sets, requiring us to add empty columns modelling 'zero occurrence' in the test data.

Table 4 shows the baseline results for both data sets. These results are computed by using the mean of the data for prediction of every instance. The columns 'training' and 'test' refer to the 75/25 split of the data set. For the last column ('ext. test'), the two previous

[21] This was done using the *caret* package in R (Kuhn 2014).

[22] All regression models were computed using the *glmnet* package in R (Friedman *et al.* 2010), which in our opinion currently offers the most transparent and flexible implementation.

[23] The procedure followed was that outlined by Nick Sabbe:
`http://stats.stackexchange.com/questions/17609/cross-validation-with-two-parameters-elastic-net-case`
– last verified: March 2018

[24] Unfortunately, we were not able to use the *glmnet* package directly to evaluate on data other than that from the training set. It seems that training and external test data would first have to be aligned in terms of features, followed by re-computation of the model and then evaluation on external test data.

Data set	RMSE		
	training	*test*	*ext. test*
two-author set	11.1	13.0	11.5
reference set	17.4	17.0	17.3/14.1

Table 4:
Baseline for both data sets

columns are added together to be used as an external validation set: i.e. the two-author model is validated on the reference data set and vice versa. There are two baselines for the reference set: the first one was calculated over the entire set, whereas the second one was based only on those items within the same time span as the two authors. Testing the two-author model on the smaller reference sample avoids extrapolation beyond the authors' time span.

5.2 *Model results*

Based on the four feature types and four n-gram lengths, sixteen different models were computed for each data set. Table 5 shows the model results for both the reference corpus (columns 2–7) and the two-author data set (columns 8–13). The first two columns for each set show the number of constant features compared to the total number of features present for each feature type and n-gram length, giving the raw counts as well as the corresponding proportions.[25] Considering these proportions with respect to feature type and sequence length (i.e. unigram, bigram, trigram, or tetragram), one can observe several patterns with respect to the number of features extracted. For both data sets, the number of all features extracted increases with n-gram size for all four feature types. However, when considering only constant features, there is a difference for the more general character and POS types as opposed to the more specific stem and lexical types. While the general types always increase in cardinality but not in proportion in the next higher sequence, e.g. unigram to bigram, across all levels, the specific types only increase up to bigram/trigram size and then decrease again. In addition, the increase in total types is considerably higher and causes the proportion of constant types of all types to be much smaller than for the first group. This is undoubtedly due to the large number of extremely rare features, adding to the count of

[25] The number of constant features reported does not include the added variables 'author' or 'year'.

Table 5: Results for the reference data set (left) and two-author data set (right) for all four feature types, showing constant features used as input versus all features extracted and the corresponding percentage in the first two column then RMSE over training and test data as well as results for testing on the other data set ('ext.test'). 'model' lists model specifications, i.e. number of β coefficients

type-ngram	Reference data set						Two-author data set					
	input		rmse			model	input		rmse			model
	constant/total	%	*training*	*test*	*ext.test*	*βs*	*constant/total*	%	*training*	*test*	*ext.test*	*βs*
Char-1	54/69	*78*	4.5	5.1	20.9	25	37/128	*29*	11.5	12.6	17.5(14.3)	2
Char-2	914/2632	*35*	3.6	5.0	80.5	24	518/3202	*16*	11.2	12.2	17.2(14.3)	2
Char-3	7236/47156	*15*	2.9	5.2	60.2	39	2788/27631	*10*	10.0	10.7	17.9(13.8)	11
Char-4	29316/350458	*8*	3.5	2.8	35.5	315	6544/137307	*5*	10.6	12.6	17.5(13.8)	5
	constant/total	%	*training*	*test*	*ext.test*	*βs*	*constant/total*	%	*training*	*test*	*ext.test*	*βs*
POS-1	43/45	*96*	4.3	4.4	21.4	10	39/45	*87*	11.6	13.1	17.5(14.2)	0
POS-2	1219/1895	*64*	3.5	3.5	18.5	83	489/1604	*30*	10.1	8.3	17.7(15.0)	69
POS-3	10973/48673	*23*	3.3	3.5	19.2	297	1461/27802	*5*	10.2	11.2	17.8(15.0)	94
POS-4	36159/593841	*0.6*	3.3	3.8	21.0	207	1547/207858	*0.7*	10.2	12.3	17.0(14.0)	1
	constant/total	%	*training*	*test*	*ext.test*	*βs*	*constant/total*	%	*training*	*test*	*ext.test*	*βs*
Stem-1	7808/320714	*2*	3.9	3.2	21.7	45	672/38915	*2*	10.2	8.8	17.8(15.1)	53
Stem-2	36189/9589629	*0.4*	3.5	3.2	12.8	81	578/967400	*0.06*	9.9	10.2	17.6(15.6)	6
Stem-3	16613/45610366	*0.04*	4.5	3.9	15.0	208	29/3079424	*0.0009*	10.4	10.2	17.3(15.0)	7
Stem-4	2238/85402502	*0.003*	5.1	5.6	15.5	55	1/4533542	*0.00002*	11.6	13.1	17.5(14.9)	0
	constant/total	%	*training*	*test*	*ext.test*	*βs*	*constant/total*	%	*training*	*test*	*ext.test*	*βs*
Lex-1	13782/741069	*2*	2.8	3.0	19.0	25	579/101630	*0.6*	10.3	9.3	18.0(14.3)	25
Lex-2	35773/11790813	*0.3*	2.9	2.2	20.6	78	633/1147540	*0.06*	10.6	10.2	17.9(14.5)	100
Lex-3	21515/47085085	*0.05*	3.4	2.4	17.0	183	78/3226493	*0.002*	11.0	9.4	17.6(16.9)	10
Lex-4	4811/89673339	*0.005*	4.3	3.4	na	35	0/4911004	na	na	na	na	na

total but not constant features. These patterns are primarily observable in the two-author data set, and are a little less pronounced for the reference data set. The remaining four columns for each set show training, test, and external test set RMSE, and the complexity of the model measured by the count of β coefficients.[26]

5.2.1 Reference corpus

We first consider models specific to the reference corpus, noting baseline results of 17.4 (training), 17.0 (test) and 11.5 (external test), as shown in Table 4. From the results in Table 5, one can observe that for character n-grams, model accuracy ranges from 2.9 to 4.5 years for the training set and from 2.8 to 5.2 years for the test set. Models 'Char-2' and 'Char-3' are best at balancing accuracy of prediction and model parsimony. With an RMSE of 20.9, 'Char-1' performs best on the two-author data, although this is still far from the baseline of 11.5, with the other three models being even less accurate (RMSE: 35–80). This suggests that there is little similarity between the data sets with regard to character n-grams. The results for the syntactic sequences (POS-n) are very regular over all four n-gram sizes, varying between an RMSE of 3.3–4.3 years for the training set and 3.5–4.4 years for the test set. External validation error on the two-author data set is lower than for the character n-grams but still not comparable with the baseline (18.5–21.4). Model complexity increases noticeably with n-gram size: our 'POS-1' model achieves an accuracy of 4.3 on the training set and 4.4 on the test set. While the bigram model 'POS-2' decreases this to 3.5 for both sets, it also adds 73 more predictors. Similarly, 'POS-3' and 'POS-4' both obtain an RMSE of 3.3 on the training set, but use 297 and 207 predictors, respectively. The word stem unigram and bigram models perform slightly better than their POS counterparts, with model accuracy slightly deteriorating after that, despite using more predictors. 'Stem-1' and 'Stem-2' achieve 3.9 and 3.5 on the training set, with 3.2 for both on the test set. This deteriorates to 4.5 and 3.9 for 'Stem-3' and then to 5.1 and 5.6 for 'Stem-4'. External validation is better than for the two previous types (12.8–21.7), but still cannot quite compete with the baseline. Overall, syntactic

[26] The coefficient count β does not include the intercept.

word features (Lex-n) and 'Lex-1', and 'Lex-2' in particular, yield the most accurate models. The unigram and bigram models obtain an error of 2.8–2.9 on the training set and 2.2–3.0 on the test set. 'Lex-1' might be considered the best model overall, as it has 53 fewer predictors than 'Lex-2', yet performs only slightly less well on the training and test sets (0.1 and 0.8 years, respectively). The external validation error (17–20.6) is higher than for stem n-grams, indicating that the two data sets might be 'closest' for that type. As previously noted, some of the above models seem rather complex and, given the tendency of elastic nets to select correlated predictors, poses the question of whether so much complexity is needed to achieve model accuracy.

In order to see which models have a large number of correlated predictors, we consider the corresponding uncorrelated models by rerunning the same experiments, but using only the lasso method, i.e. setting α to 1. This highlights several aspects of the regression models computed earlier: a simple model of ~10–30 predictors can still be improved by adding features, in the sense that these contribute enough new information to improve prediction accuracy. In most cases, however, adding more features to a model of 80 predictors rarely improves prediction accuracy. Compare adding 7 features to achieve a −0.3/−0.5 error decrease ('Lex-4') to adding 151 features for a −0.5/−0.5 RMSE decrease ('Lex-3') for training set and test set respectively. What is also notable is that most lower n-gram models do not have any correlated predictors, seeing that elastic net and lasso methods yield the same models, whereas the number of correlated predictors rises with n-gram size up to trigram size, whereafter model size suddenly decreases more or less dramatically.[27] This strongly suggests that there is most overlap for trigram models on the most changing features used in each time slice. Thus, while there is likely to be most background language change in syntactic word features, all types produce accurate enough models to suggest that reasonably interesting temporal change must have taken place. The language change aspect is examined in more detail in Section 6.

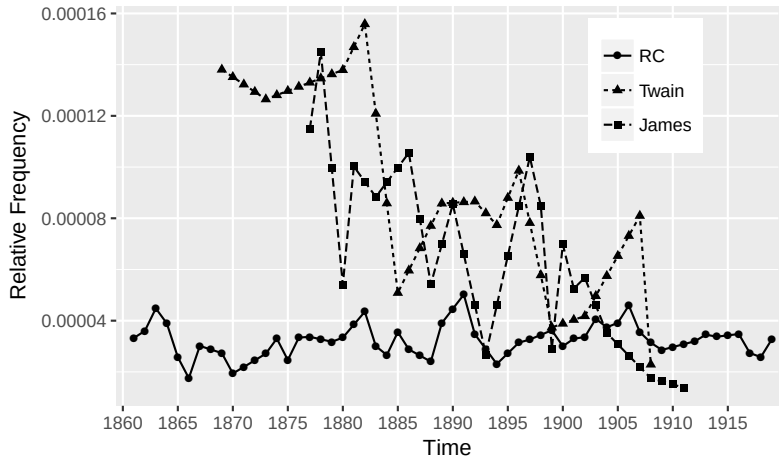
[27] This is with the exception of character n-grams, as these would probably need to grow to average word length in order to be less correlated.

5.2.2 Two-author data set

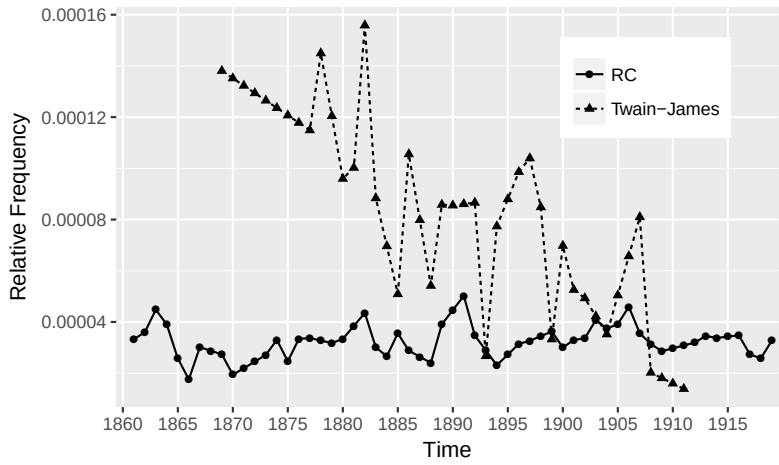
We now turn to the models intended to capture individual change, specifically in James' or Twain's language. The baseline results for the two authors yielded 11.1 (training), 13.0 (test) and 17.3/14.1 (external test). Beginning with the character n-gram models, Table 5 shows that 'Char-1' and 'Char-2' are very close to the baseline, containing very few predictors, indicating that these two types carried little discriminatory power. The trigram model 'Char-3' is the best character model, with 10/10.7 RMSE for training and test set, where the error is much lower than the baseline of 13, especially for the test set. The 'Char-4' model does not quite reach the same accuracy, although it is an improvement on the first two models. The results on the external test data are consistently congruent with the baseline for that set. Moving on to syntactic sequences, the unigram model 'POS-1' is actually the null model, as it is the most parsimonious model within one standard error of the best model with 38 features, suggesting that this type is not discriminatory enough in relation to publication year. The best POS model is 'POS-2' with 10.2/8.3 on training and test set respectively, but it increases complexity by adding 69 predictors. 'POS-3' adds even more complexity (94 predictors), but performs worse than 'POS-2'. Interestingly, the 94 predictors in 'POS-3' have the same predictive power on the training set as 'POS-4's one and only predictor ⟨VBD VBN IN JJ⟩.[28]

Figure 1 depicts the tetragram ⟨VBD VBN IN JJ⟩ for Twain and James individually (Figure 1a) and combined together (Figure 1b). Even though relative frequency values vary over only a small range (0.00004–0.00016) for both James and Twain (Figure 1a), there is a discernible downward trend over time, offering a fair indication of temporal origin. The combined plot, though a generalization, still presents a fair approximation of each individual plot. In comparison, the same feature exhibits less of a trend over time for the reference corpus. The prediction accuracy of stem models is comparable to that of character and POS n-grams, while models tend to be more parsimonious. Results range from 9.9–10.4 on the training set and 8.8–

[28] This tag represents a sequence of ⟨a verb in past tense (VBD), a verb in past participle (VBN), a preposition (IN) and an adjective (JJ)⟩ as in ⟨were.VBD accompanied.VBN by.IN restless.JJ⟩.



(a) ⟨VBD VBN IN JJ⟩ for James, Twain, and the RC

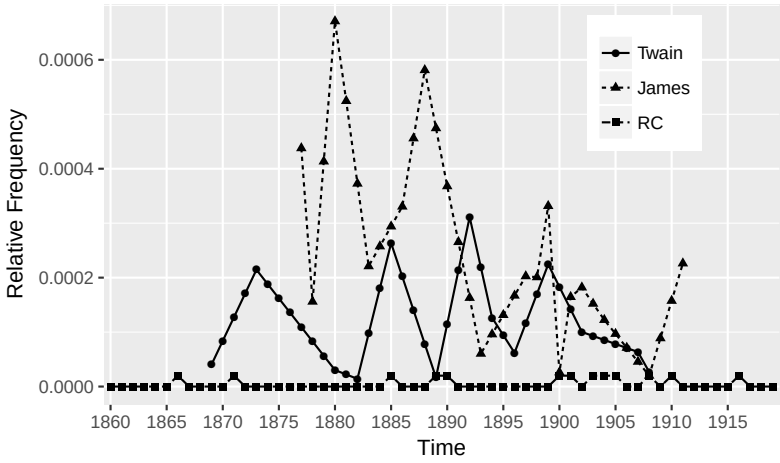


(b) ⟨VBD VBN IN JJ⟩ for James + Twain, and the RC

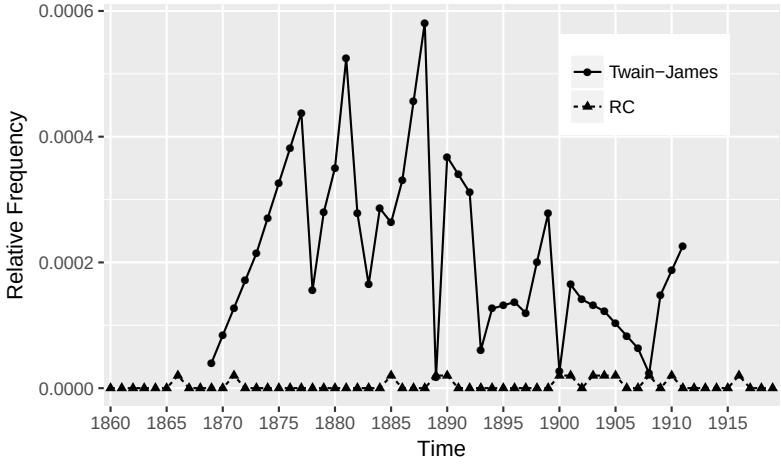
Figure 1: The development over time of the tetragram ⟨VBD VBN IN JJ⟩, showing relative frequency in relation to all tokens for the reference corpus (RC) and for the two authors. Figure (a) shows the feature for Twain and James separately. Figure (b) shows the combined two-author frequency, averaged for years when both published work

10.2 on the test set for 'Stem-1', 'Stem-2' and 'Stem-3'. For word stem tetragrams, the number of constant features drops to one (which is the feature ⟨i don t know⟩), causing the null model to be selected.[29] Figure 2 depicts this feature for Twain and James separately (Figure 2a) and combined into one (Figure 2b), each time alongside the reference corpus. Variability somewhat decreases over time for the two

[29] The corresponding syntactic word feature would be: ⟨i do n't know⟩.



(a) ⟨i don t know⟩ for James, Twain, and the RC



(b) ⟨i don t know⟩ for James + Twain, and the RC

Figure 2: The stem feature ⟨i don t know⟩ for the reference corpus, and for Twain and James separately (Figure (a)) and combined (Figure (b))

authors, if less markedly than in the previous case, and while there is a downward trend for James, there is no specific trend visible for Twain. Combining their two plots over time yields a less appropriate approximation to each individual, indicating that there are stronger differences between them. Interestingly, this tetragram feature was not constant over the reference corpus, in spite of a much larger data selection available – its line in the plots indicates occurrence rather than relative frequency in both Figures 2a and 2b. When the feature

occurs, the raw count generally varies between 1 and 2 and never exceeds 6 (total token count for the same year is 2,228,655). This indicates a very different usage from James and Twain, and could imply that other synonymous forms were more common, e.g. 'I do not know' or that first person references were used less frequently than by the two authors. Examining alternative, high-ranking models for 'Stem-4' yields a pairing of ⟨i don t know⟩ with the 'author' feature. Figure 2 shows that relative frequencies for James and Twain are reasonably different until 1890, with little overlap, possibly rendering separation by authorial source more useful than in the previous cases.

This result shows that, although this feature was used by both James and Twain, it was rare in general language at the time. James initially used it more than Twain, but, over time, their rates of use appear closer. Thus, there are two different dimensions to this analysis, the constancy of a feature over a corpus, and its relative frequency. The main difference between the reference corpus and the two-author data set is that of constancy, whereas the main difference between Twain and James pertains to the feature's relative frequency. In any case, a more detailed investigation is needed to exclude possible confounding factors, such as genre or narrative perspective, to confirm that this pattern is rooted in stylistic differences only.

Finally, we consider the most specific linguistic type, syntactic word features. The best overall models are 'Lex-1' and 'Lex-3', with 10.3/11 on the training set and 9.3/9.4 on the test set. 'Lex-2' is more complex (100 predictors) and yet a little less accurate.

These results suggest that the more general feature types (character/POS) need longer sequences to be discriminative. In contrast, stem n-grams are fairly accurate, sometimes even with only very few predictors, provided there are enough input features. The fact that the 'author' variable was never chosen to be a part of any model suggests either that Twain and James are rather similar with respect to their shared constant features that are discriminatory over time, or that their rate of change is entirely different, making a distinction for the level not helpful.

5.3 *Comparison with previous results*

The final part of the experiments is to compare these results with those from our previous study on syntactic word unigrams (Klauss-

Table 6:
Results for previous work (Klaussner and Vogel 2015), showing RMSE and model size for the reference corpus, the James and Twain data set, and the combination of all three data sets

Reference set				
Model	*training*	*test*	*ext.test*	βs
1	3.2	4	15.4(T)/20.3(J)	4
2	11.9	12.1	42.2(T)/44.7(J)	5
Two-author set				
Model	*training*	*test*	*ext.test*	βs
1	5.5	7.2	–	5
2	5.2	8	–	7
Combined set				
Model	*training*	*test*	*ext.test*	βs
1	2.8	1.8	–	5

ner and Vogel 2015). Table 6 shows the results for the reference corpus, the two-author data set, and a third corpus combining all data sets in one. In comparison to earlier experiments, our results for the reference corpus add ~1 year accuracy in prediction. The results for the two-author data set are less accurate. This confirms that taking only constant features for prediction and discarding all others results in the loss of valuable predictors. In part this could be due to a feature's non-occurrence in particular years, possibly aiding the statistical technique to discriminate more easily between years. Using features occurring less reliably has to be applied with caution as, on the very infrequent side of the frequency spectrum, there lurks statistical optimization, which would not only yield unstable models, but would also focus less on characteristic and more on idiosyncratic aspects of the particular data set under study. One therefore needs to differentiate between features that are infrequent during an author's lifetime, but very frequent in those years when they do occur, and features that are consistently infrequent. An extreme case of this would be sets of *hapax legomena*. The reason why the models are more accurate for frequent, but not quite constant, features may be that authors are likely to be more consistent for features that they use constantly throughout their literary career, than for those that they use less regularly. In any case, we emphasize that our purpose is not achieving the highest possible accuracy in assignment of temporal provenance, but in understanding what fea-

tures change in frequency over time, and how those changes are to be interpreted. The latter task is open-ended, but depends on the former.

6 ANALYSIS OF LANGUAGE CHANGE

In this section, we consider salient features of the regression models presented in Section 5.2. In order to select those features that change most over time, we rank the respective model's predictors according to the absolute weight it is assigned in the model, thereby selecting features that increase and decrease linearly over time. However, to identify features that did not exhibit any change over time, we had to exclude features that rated high on either linear or non-linear change. For this purpose, we evaluated all features separately with respect to the response variable, and selected those that rated low on both linear and non-linear relationships. Section 6.1 introduces some general language change trends and Section 6.2 then analyses the data for the two authors in comparison with the reference corpus.

6.1 *Reference language change*

In the following, we present some aspects of general language change based on the changes detected in the reference corpus. This is not presented as an exhaustive list, but merely as a series of examples. In the following, we focus on lexical and syntactic change.

Figure 3 shows samples of the highest-rated features for each of the three categories: 'increase over time', 'decrease over time' and 'no change'. Considering shorter n-gram sizes shows that there might be considerable overlap between different models of the same feature type but different n-gram size, and also between different feature types. Figure 4 shows the word n-gram ⟨a matter of fact⟩ and its hypergram ⟨a matter of⟩. As can be seen from the difference in frequency, there are a number of other frequent realizations of ⟨a matter of⟩, such as ⟨a matter of concern⟩ or ⟨a matter of urgency⟩. There are cases where the more specific sequence accounts for most of the occurrences of the generic one, whereas in cases like these it only accounts for part of them.

Figure 5 shows the most prominent syntactic tetragrams. The sequences ⟨DT NN IN WRB⟩ and ⟨DT NN TO VBG⟩ both increase over

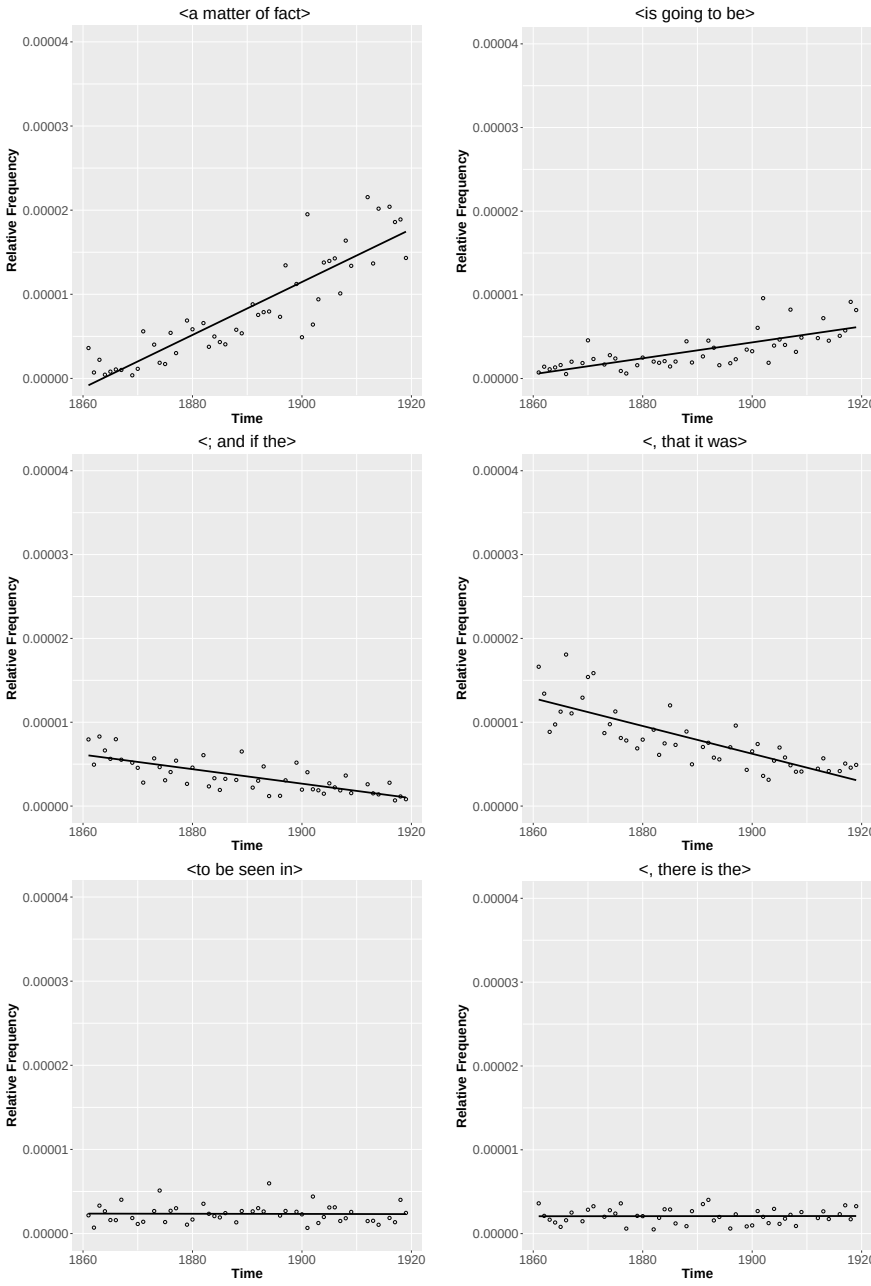


Figure 3: Reference corpus: relative frequency of several syntactic word tetragrams, exhibiting ‘increase’, ‘decrease’, or ‘no perceptible change’ over time

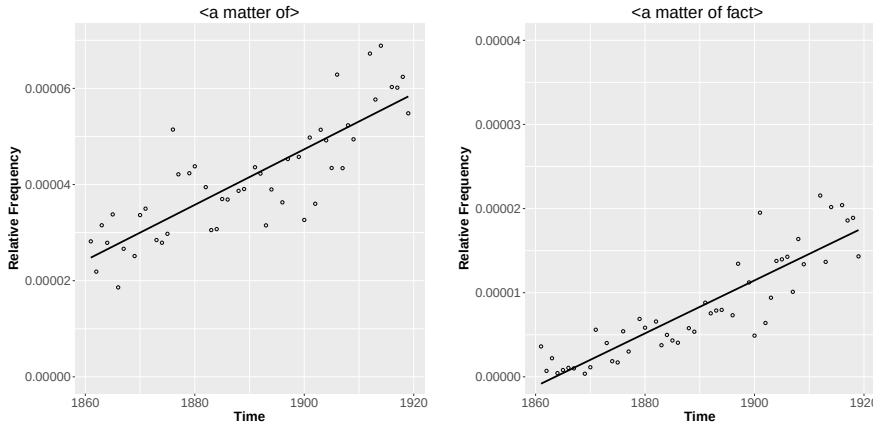


Figure 4: Reference corpus: relative frequency for ⟨a matter of⟩ and ⟨a matter of fact⟩

time. Phrases such as ⟨the fact that when⟩ or ⟨the secret of where⟩ are examples of the former, and ⟨no objection to saying/taking⟩ or ⟨a view to showing/discovering⟩ are examples of the latter. Thus, depending on whether the words in the sequence are content or function words, and whether they are part of a collocation, certain combinations will be more frequent (⟨a view to⟩/⟨no objection to⟩), while others may be more variable. The shorter variant of this ⟨DT NN TO⟩ does not seem to be discriminative over time. Similarly, examining some corresponding syntactic word sequences ⟨a.DT view.NN to.TO⟩ and ⟨no.DT objection.NN to.TO⟩ shows that, although constant, they do appear to change in a rather random fashion. The more specific tetragram sequences, such as ⟨no objection to saying⟩ are usually not constant. Realizations of decreasing POS features (⟨CC NN VBP PP⟩ and ⟨IN VBG , IN⟩), also yield patterns of fixed and varying units: ⟨and pride/happiness attend her⟩ and ⟨by saying, that⟩/⟨without murmuring, because⟩. The syntactic combinations that show the least development during this time span are ⟨EX VBZ RB JJR⟩ with examples such as ⟨there is far more/less⟩/⟨there's something stronger⟩, and ⟨VBD NN DT NN⟩ with examples like ⟨was nothing the matter⟩ or ⟨made music all day⟩.

Given the size of the corpus, one would expect a variety of feature realizations to be among the constant features, especially in the presence of multiple genres, and the differences in language us-

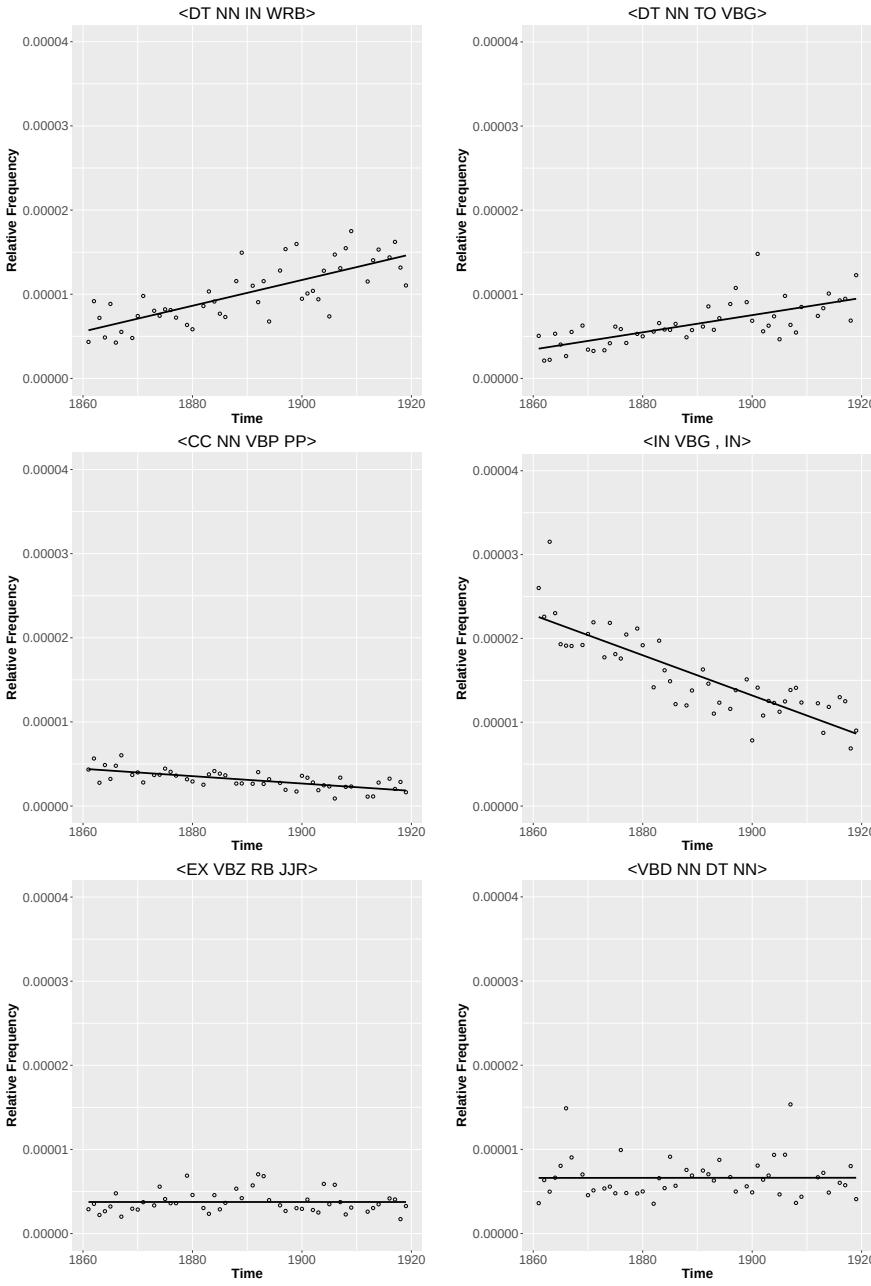


Figure 5: Reference corpus: relative frequency of several syntactic tetragrams, exhibiting ‘increase’, ‘decrease’, or ‘no perceptible change’ over time

age found in these genres. In spite of this, most of the consistent features or their generalizations present here seem to be expressing opinions, or to be ways of organizing these, such as ⟨a matter of fact⟩ or ⟨a view to⟩/⟨no objection to⟩, which are items that could be expected to appear in a variety of contexts. In order to identify change that is not general to all written language, one might investigate change in different genres, such as fiction, or newspaper articles. The most dramatic change is found in very general POS n-grams, which incidentally also display more spread. In contrast to syntactic word n-grams, POS n-grams are more volatile in that they represent a group of words that could possibly change or give rise to different frequencies.

6.2 *Two-author language change*

We now turn to the analysis of the two authors, to examine how their language changed or stayed the same over time, while also taking into consideration how their language differed from the reference language of the time. In the following, we consider different aspects of how style could vary. Section 6.2.1 considers differences between constant feature sets of lexical types. Sections 6.2.2 and 6.2.3 consider stylistic differences between the reference corpus and the two authors, and then any stylistic differences between the two authors.

6.2.1 Constant features

In order to explore the stylistic differences between Mark Twain and Henry James, we examine different sets of constant terms: those they share and those they do not share. It is important to note that constancy does not necessarily imply high frequency, and that one word or expression could be constant for only one author but more frequent overall for the other.

Figure 6 shows 'wordclouds' based on their individual non-shared noun, interrogative pronoun, and adjective type features. We grouped these together for inspection since they could all occur in noun phrases but, unlike pronouns and determiners, are less grammatically controlled, and therefore more meaningful.

Table 7 shows the relative frequency data for wordcloud items, ordered by relative frequency, showing the median rank of each item in the wordcloud group, and among all constant features for that author.

Table 7: Relative frequencies and rank for the 20 most frequent wordcloud items, in the works of Twain, of James, and of both authors together. Words are listed by relative frequency rank (RFR). The 'wcr' columns show wordcloud group ranking. The 'cr' columns show rank among all constant items

RFR	Individual items								Shared items							
	Twain	*RF*	*wcr*	*cr*	*James*	*RF*	*wcr*	cr	*Twain*	*RF*	*wcr*	*cr*	*James*	*RF*	*wcr*	*cr*
1	god	0.00039	10	344	mr	0.00162	1	112	what	0.00239	1	63	what	0.00383	1	38
2	money	0.00037	1	266	lady	0.00053	2	206	time	0.00184	2	72	little	0.00177	2	65
3	boy	0.00037	4	316	de	0.00047	168	722	man	0.00145	3	79	who	0.00146	4	76
4	mother	0.00033	7	348	father	0.00046	23	399	other	0.0014	5	78	time	0.00134	3	80
5	everybody	0.00031	2	264	whom	0.00045	3	207	good	0.00132	4	87	great	0.00119	7	89
6	body	0.00029	3	276	lord	0.00040	134	733	who	0.0013	6	80	other	0.00117	5	102
7	sir	0.00028	19	381	dear	0.00034	4	260	way	0.00129	7	88	young	0.00107	14	117
8	boys	0.00028	17	385	companion	0.00034	5	242	old	0.00113	11	104	way	0.00104	9	103
9	dead	0.00027	6	275	charming	0.00029	17	352	little	0.00112	10	103	moment	0.00104	6	99
10	door	0.00027	5	319	effect	0.00028	7	273	thing	0.00105	9	106	man	0.00101	8	101
11	family	0.00027	9	311	particular	0.00027	10	305	day	0.00097	8	100	good	0.001	10	110
12	children	0.00024	14	330	view	0.00027	12	334	people	0.00096	12	116	nothing	0.00099	11	105
13	ready	0.00023	8	333	possible	0.00026	8	287	great	0.00082	13	121	more	0.00095	12	114
14	anybody	0.00023	12	409	reason	0.00026	9	318	nothing	0.00079	14	134	things	0.00094	15	118
15	nobody	0.00021	13	375	round	0.00025	6	291	more	0.00073	15	136	own	0.00091	16	123
16	week	0.00021	16	400	impression	0.00024	15	327	place	0.00067	18	158	old	0.00088	19	127
17	river	0.00021	22	426	tone	0.00023	13	321	last	0.00067	16	163	something	0.00082	13	119
18	girl	0.00021	20	441	rate	0.00023	11	319	things	0.00065	19	159	thing	0.00081	20	131
19	minutes	0.0002	23	447	love	0.00022	25	396	years	0.00064	17	150	last	0.00076	18	130
20	bed	0.0002	18	436	bad	0.00022	14	322	night	0.00063	20	155	eyes	0.00074	17	128

Figure 6: Noun, interrogative pronoun, and adjective type wordclouds for Twain (left) and James (right), based on non-shared constant features

Twain's most prominent words express existential concepts, apparently pertaining to a more questioning nature, e.g. 'god', 'money', ' everybody', 'anybody', 'nobody', 'family', 'mother', 'children', 'dead', 'heaven', 'church', 'trial', and 'soul'. In contrast, James' most prominent words in this group are more prosaic, e.g. 'mr', 'father', 'lady', 'dear', 'whom', 'lord', 'charming', 'companion', 'impression'.[30] It is interesting to note the difference between James' most frequently used form of address, 'Mr', and Twain's 'Sir' – 'Mr' suggests that one could address both a superior and an equal, whereas 'Sir' is used predominantly when addressing a superior, which is plausible as Twain also wrote about less wealthy people.[31] James' list also includes the French word 'de', often found in names and addresses and, which was incorrectly tagged here as a proper noun.[32] There are some other interesting contrasts, such as 'conscience', which is constant for Twain, and 'conscious'/'consciousness', constant for James. Twain's words suggest more intense situations, intimating both good and bad, e.g. 'crime', 'cruel', 'blood', 'dark', 'lonely', 'alive', 'peace'. James' most negative words in this group are 'sad', 'helpless', 'victim', indicating that Twain's language was more explicit. While James' stories do contain conflicts, they were possibly more veiled than in Twain's texts.

[30] As all data was transformed to lowercase for analysis, words, such as 'Mr' appear that way in figures as well.

[31] The word 'Sir' is ranked 19 among wordcloud features and 381 among Twain's constant features.

[32] The word 'de' is ranked 168 among wordcloud features and 722 among James' constant features.

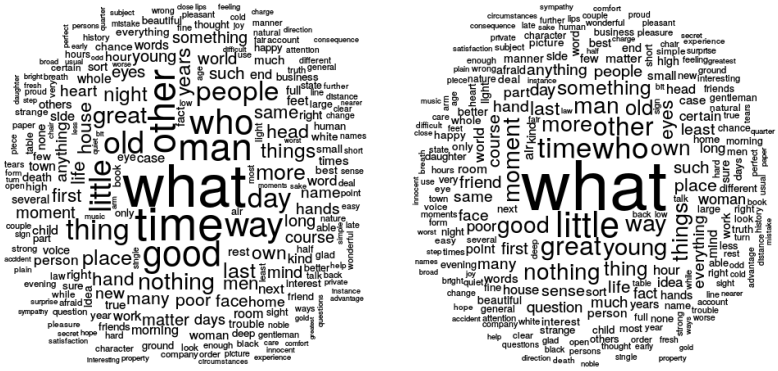


Figure 7: Noun and adjective wordclouds for Twain (left) and James (right), based on their shared constant features

Figure 7 shows the wordclouds for their shared constant nouns, interrogative pronouns, and adjectives. Their most prominent words are quite similar here, e.g. ‘what’, ‘time’, ‘little’, ‘good’, and ‘young’. There are some less frequent words for both that are interesting to consider, with a wider semantic range: ‘circumstances’, ‘feeling’, ‘consequence’, ‘believe’, ‘truth’, and ‘pleasure’. Depending on context, these words might take on either a more superficial or deeper meaning, e.g. ‘I believe you’re right’ and ‘I believe in one Christ’.

Interestingly, both authors took an avid interest in history, evidenced by the syntactic unigram ⟨history⟩ being among their shared constant features. Both Blair (1963) and Thomas M. Walsh and Thomas D. Zlatic (1981) note that history played an important part in Twain’s personal as well as his professional life, even if he did not always incorporate his knowledge consistently into his works (Williams 1965). In his 1884 essay ‘The Art of Fiction’, James actually claims his place among historians, since a novelist chronicles life, and as ‘picture is reality, so the novel is history’ (James 1884). All of the two authors’ constant word unigrams are present in the constant features of the reference corpus, except for James’ term ‘vagueness’.[33]

While constant word unigrams reveal a great deal about recurring concepts, longer sequences might hold more information about unique aspects of style, as these tend to be more generic. Table 8 shows examples of constant bigram and trigram word sequences and

[33] Although using wordclouds can give some insight into the data, it cannot replace the study of actual word frequency distributions. The extended set of constant features can be found here: `www.scss.tcd.ie/clg/4thIWCH/`.

Table 8: Examples of constant syntactic word sequences (bigrams and trigrams) characteristic of: both authors (columns 3–7); James alone (columns 8–10); Twain alone (columns 11–13). The column 'cr' indicates median rank, considering bigram and trigrams separately. For readability, relative frequencies are multiplied by 100. In the interest of space, ⟨...⟩ are omitted here

Group	Type	shared bigrams/trigrams					James only			Twain only		
		n-gram	*RF(J)*	*cr(J)*	*RF(T)*	*cr(T)*	*n-gram*	*RF*	*cr*	*n-gram*	*RF*	*cr*
Possessives	n2-IN-m	by his	0.003	589	0.005	589	as his	0.005	795	into his	0.012	1069
	n2-IN-f	with her	0.007	130	0.005	547	on her	0.029	301	–	–	–
Body parts	n2-sg	hand ,	0.004	428	0.011	515	eye ,	0.025	1116	head ,	0.093	665
	n2-pl	eyes ,	0.009	446	0.046	815	hands ,	0.051	1016	–	–	–
	n2-PP$-m	his eyes	0.003	414	0.003	822	–	–	–	his eye	0.035	995
	n2-PP$-f	–	–	–	–	–	her head	0.039	431	her face	0.01	824
Expressions	n3	a matter of	0.004	158	0.046	284	in spite of	0.016	24	on account of	0.007	169
Consciousness	n2	i know	0.008	198	0.007	225	i mean	0.06	258	i knew	0.008	569
	n3	, i think	0.012	185	0.008	307	i think ,	0.005	154	. i know	0.007	233
Existential	n2	there is	0.035	296	0.048	157	there are	0.01	467	there 's	0.004	487
	n3	. there was	0.004	40	0.004	33	–	–		, and there	0.004	84

their frequencies found in the data for Twain, for James, and for Twain and James together. These lists are mutually exclusive, meaning that each term is shown only once, in the set where it is most constantly used. The rows group together n-grams by selection category. The first group contains bigram sequences of a noun followed by a preposition followed by either a male or female possessive pronoun. The second group contains singular or plural body references, either followed by a comma, or preceded by a male or female possessive pronoun. The third group contains expressions that are used for emphasis or contrast. The last two groups focus on items expressing some epistemic commitment, or with an existential construction.

Twain's language, in particular, abounds with a great variety of body references, some of which are also used by James. However, James tends to focus on body descriptions, e.g. 'face', 'eyes', 'hands', whereas Twain's constant terms include items used more abstractly, such as 'heart'. Twain's language also features many more 'existential' constructions, such as ⟨there 's⟩, which are also found in James, but with less variety. Both authors use expressions indicating reflection or thought (⟨I know⟩, ⟨I think⟩, etc.). Twain's constant terms also include the expression ⟨don't know⟩, which James does not appear to use. James seems to use contrasting features more often, e.g. ⟨in spite of⟩ or ⟨, however ,⟩, which Twain appears to employ more sparingly. Both use the male perspective more than the female one, i.e. their constant feature lists both contain various possessive and regular pronoun constructions for male characters, which are not present in the same quantity for female characters.

However, in order to properly verify these impressions, one needs to take a closer look at the actual number of constructions in each group, and their respective frequencies. We begin by considering constructions containing existential 'there' and its overall unigram relative frequency in all three corpora; the corresponding plot is shown in Figure 8. On average, Twain's usage is a little higher (ca. 0.002) than that of James and of the reference corpus, which are both around 0.0018. Table 9 shows details about the number of types for a particular item, for instance in what constructions the feature ⟨there.EX⟩ appears. This shows that Twain clearly has more constant existential types than James and, as the frequency analysis showed, he also uses

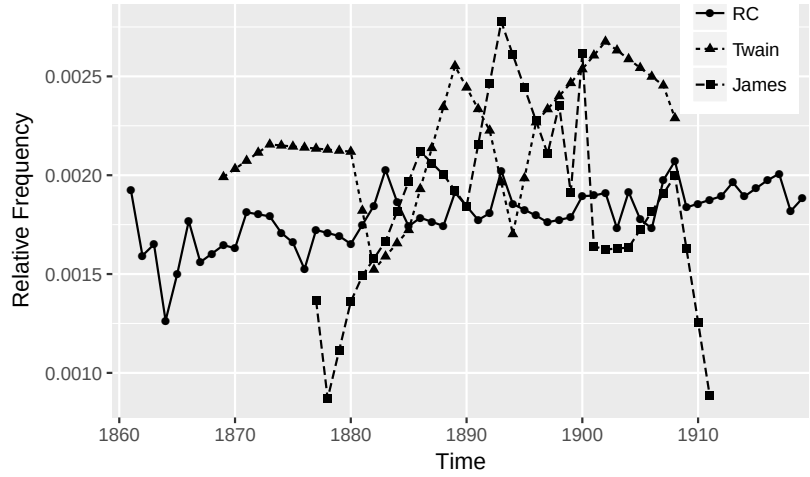


Figure 8: Existential ⟨there⟩ for all three corpora

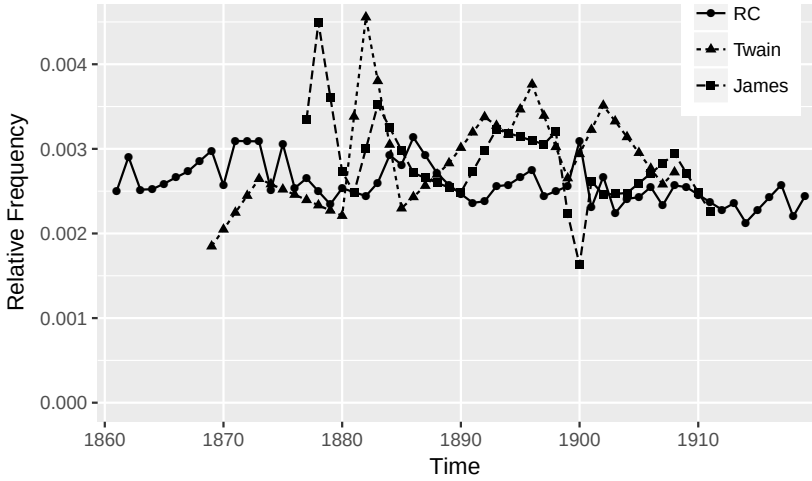
Table 9: Frequency and number of feature types for prominent constructions

	James			Twain			Shared	
Variable	*1-gram:μ*	*Lex2*	*Lex3*	*1-gram:μ*	*Lex2*	*Lex3*	*Lex2*	*Lex3*
⟨*there.EX*⟩	0.0018	3	–	0.002	7	4	4	2
body parts sing	0.0028	12	4	0.0029	14	2	4	–
body parts pl	0.001	1	–	0.001	5	1	3	–
female pr	0.023	8	1	0.008	20	5	17	–
male pr	0.025	9	2	0.025	49	54	27	9

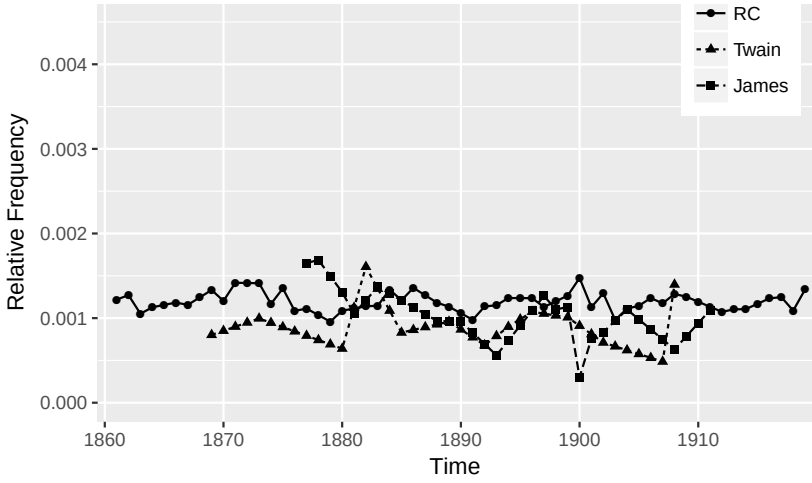
them more often. There is also an increase in usage over time for both authors, as well as for the reference corpus.

Figure 9 depicts frequency rates for body references: Figures 9a and 9b show singular and plural body parts, respectively. Interestingly, average use for body references lies above the reference corpus for singular items and below it for plural items, in both Twain and James.[34] There seems to be a decrease in usage for both types over time, with a more dramatic decrease for plural body parts. The difference between the two authors lies primarily in the variety of constructions used: there tends to be more variety in Twain's constant features – this does not mean that James does not use these features at all, but that there are fewer features that James uses regularly.

[34] The frequency rates for the reference set are 0.0026 and 0.0012, respectively.



(a) Frequency of singular body parts for James, Twain, and the RC

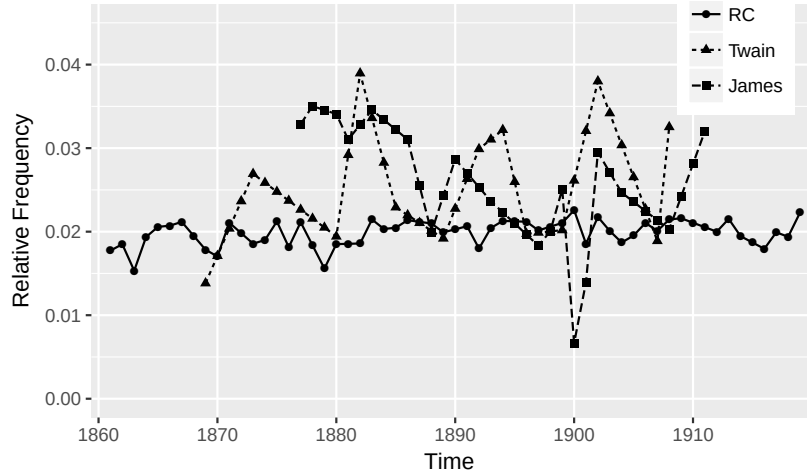


(b) Frequency of plural body parts for James, Twain, and the RC

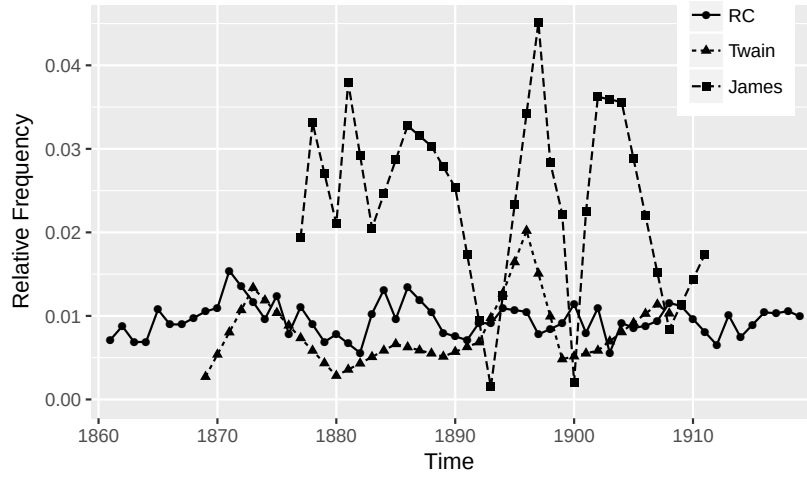
Figure 9: Body part constructions for all three corpora

Figure 10 shows the frequency rates for possessive and regular pronouns, with masculine forms in Figure 10a and feminine forms in Figure 10b. Both authors use the male perspective much more than was usual for the time, compared with the average rate of 0.025 to 0.02 in the reference corpus. Furthermore, James (0.023) refers to women through female pronouns more than twice as much as Twain (0.008), or the reference corpus (0.009). Incidentally James' constant bigram list also includes ⟨woman ,⟩ and ⟨women ,⟩ – it thus appears

Figure 10: Male and female references for all three corpora



(a) Frequency of male pronouns for James, Twain, and the RC



(b) Frequency of female pronouns for James, Twain, and the RC

as though James focused his narrative on women much more than was usual for his time. In contrast, Twain has markedly more varied constant constructions featuring pronoun references, especially for males. This could mean one of two things: either that he is more variable in his language describing people, given that he has more common phrases, or in fact that he is less variable, as he tends to draw more often from a limited set. Without a comparison with more contem-

poraneous authors, to examine the proportion of gendered pronoun constructions in their non-constant bigrams, it is not clear whether this aspect is usual or unusual. For instance, James might only have a few constant constructions, changing his language use depending on the situation.

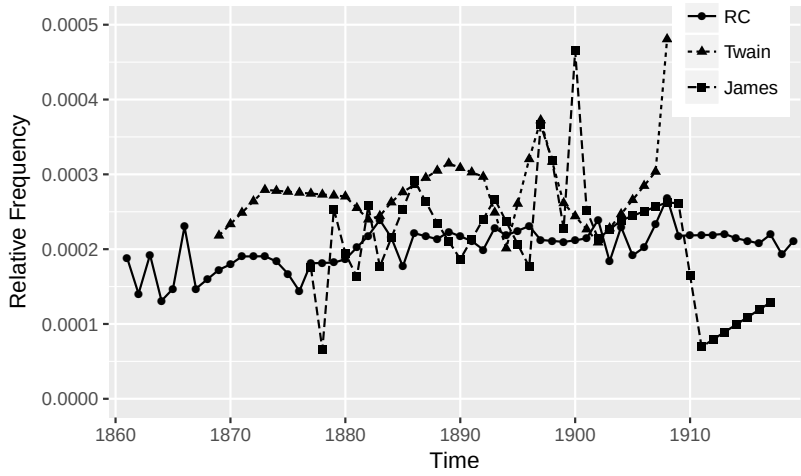
6.2.2 Stylistic differences with the reference set

In order to explore any differences from the reference language, we consider the shared salient features, i.e. the features that appear in Twain, in James, and in the reference corpus. Among the character n-gram models, there are no common predictors, except for the letter ⟨q⟩ in the unigram model. All models have a positive weight for this predictor, but only the authors show a clear upward trend over time. All word stem and syntactic word n-gram models yield one shared bigram ⟨, by⟩, which is shown in Figure 11 and Figure 12, together with three highly weighted shared POS bigrams ⟨CC EX⟩, ⟨WDT ,⟩ and ⟨MD ,⟩.

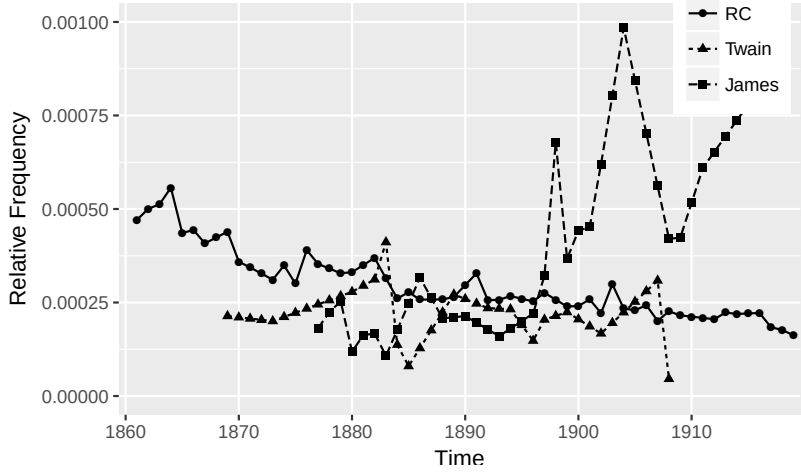
The bigram ⟨CC EX⟩ realizes expressions such as ⟨but there⟩ or ⟨and there⟩, that have already been mentioned earlier with respect to the constant features in James and Twain. For this POS bigram, their average rate tends to be higher than that of the reference corpus. What is noticeable for the other three features is that the three data sets are rather well separated, with James having the highest usage of all. This will be explored in more depth as part of the between-author analysis in Section 6.2.3. All lines show some development over time, explaining why these are salient features in the models.

With respect to syntactic changes, there seems to be a marked reduction in noun phrase constructions for the two authors, a reduction that is not present in the reference corpus, as shown for two examples in Figure 13. This trend can also be observed in several other noun-phrase-based sequences, such as ⟨DT NN NN⟩, ⟨IN NN NNS⟩, ⟨JJ NN NNS⟩, ⟨NN NNS⟩, and ⟨NN NNS SENT⟩. Examining general counts over all unigram noun-related POS tags, i.e. ⟨NN⟩, ⟨NNS⟩, ⟨NP⟩, ⟨NPS⟩, returns somewhat inconclusive results. Both authors show a decrease for ⟨NNS⟩, and Twain also for ⟨NPS⟩. In contrast, there is a slight increase in pronouns in Twain's data. Overall, this might indicate a shift in how noun phrases are commonly constructed, i.e. simpler or more pronoun-based. Merely summing the tags does not

Figure 11: Prominent features common to Twain, James, and the reference corpus

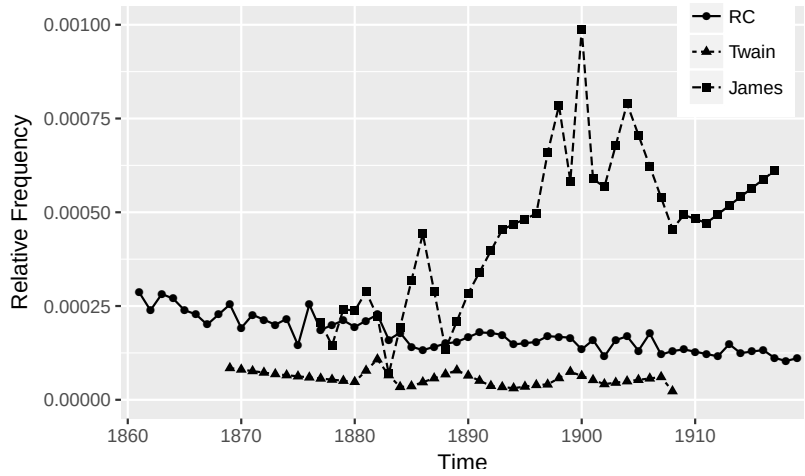


(a) Frequency of ⟨CC EX⟩ for James, Twain, and the RC

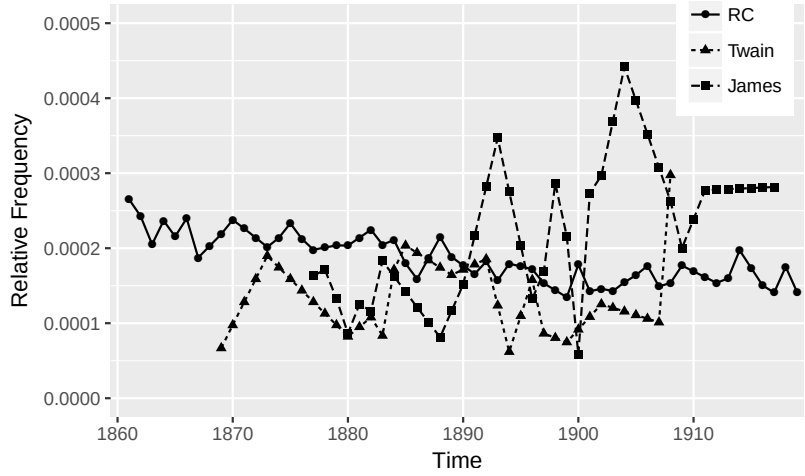


(b) Frequency of ⟨, by⟩ for James, Twain, and the RC

adequately describe how many noun phrases there are, nor how they are composed. Nor would simply looking at a rise or decrease in determiners suffice to ascertain how the above items are distributed. This result can only provide pointers for interesting aspects to consider in future work, which would require actually looking at the number of noun phrases overall and investigating whether the way they are composed changes over time.



(a) Frequency of ⟨WDT ,⟩ for James, Twain, and the RC



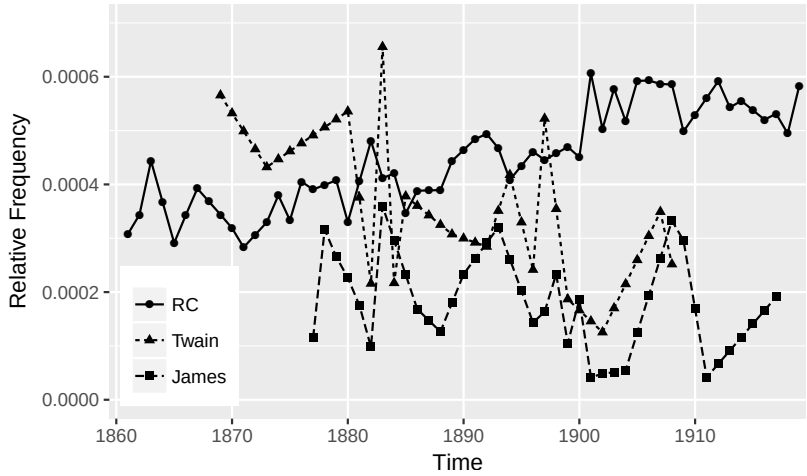
(b) Frequency of ⟨MD ,⟩ for James, Twain, and the RC

Figure 12: Prominent features common to Twain, James, and the reference corpus

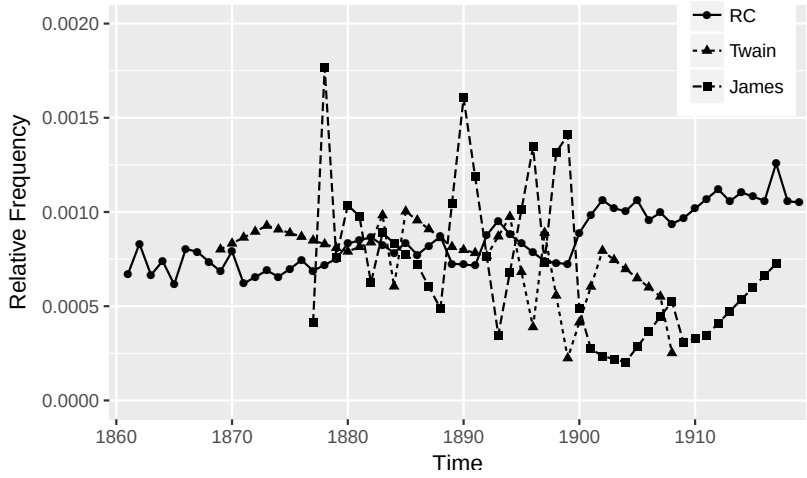
6.2.3 Stylistic differences between authors

In this final part, we consider some stylistic differences between the two authors. Although the graphs in Figure 11 have already been discussed as part of the comparison between reference corpus and author-specific models, these features are also interesting to analyse in terms of what this difference in usage implies about differences in authorial style. Three of these features (⟨WDT ,⟩, ⟨MD ,⟩, ⟨, by⟩) are certainly

Figure 13: Decrease in two noun phrase types for the two-author data set and increase for the reference set



(a) Frequency of ⟨DT NN NNS⟩ for James, Twain, and the RC



(b) Frequency of ⟨NN NN SENT⟩ for James, Twain, and the RC

more important for James, as Twain's usage mostly lies below that in the reference corpus. Examining some of the lexical realizations for these features for James and Twain shows clear differences in usage. James seems to use these features to build longer and more complicated sentences, increasing the cognitive workload on the part of the reader, which probably contributed to James' later style being considered somewhat 'obscure' and 'over-planned' (Beach 1918); an example of ⟨WDT ,⟩ is shown in (4).

(4) *It sounds, no doubt, too penetrating, but it was by no means all through Sir Claude's betrayals that Maisie was able to piece together the beauty of the special influence through which, for such stretches of time, he had refined upon propriety by keeping so far as as possible his sentimental interests distinct.*

There are a few instances of simpler constructions, not introducing a proper sub-clause, such as 'of which, however, she had', but these examples appear to be less numerous overall. While Twain's texts do contain these types of constructions, they appear more sparingly and also take a different, less convoluted shape, an example of which is shown in (5).

(5) *There is only a plausible resemblance, which, while it is apt enough to mislead the ignorant, cannot deceive parties who have contemplated both tribes.*

(6) *Then it is, in the final situation, that we get, by a backward reference or action, the real logic and process of the ambassador's view of how it has seemed best to take the thing, and what it...*

(7) *Without suspecting it, Dr. Peake, by entering the place, had reminded me of the talk of three years before.*

Examples of the syntactic word bigram ⟨ , by⟩ are shown in (6) and (7) for James and Twain, respectively.

7 DISCUSSION

This work has presented various experiments and analyses aimed at discovering salient features of general and individual language change. To identify these features, we used linear regression models, retaining only constant features for the reference corpus models, and shared constant features for the two-author models. Selecting only constant features serves to focus the analysis on the features the authors remained true to over their creative life span. Features used in a non-constant fashion would be interesting to analyse to complement the current results. We chose to only use linear models, for our experiments here, to limit the quantity of results. Other types of models should be studied in future work. As we chose to consider different feature types and n-gram sizes, there were many results and interpre-

tations to consider, and unfortunately we could not do justice to them all. The interpretations that we have provided are subjective, yet anchored in the critical literature that we have explored to date. We hope that other researchers will identify other natural categories within the features marked as salient by our methods, which may support competing interpretations. Our task in this work is not to propose definitive interpretations, but to provide methods to highlight features that undergo interesting development during writers' careers and to suggest that these interpretations may be anchored in critical responses to the career.

In terms of general differences from the reference corpus, there seems to be an interesting shift for both authors towards the use of simpler noun phrase constructions. We could not clearly identify all the particulars as part of this work. It would probably not suffice to simply analyse the composition of noun phrases, as genre and authorship could play a factor in this as well. One would therefore need to consider other contemporaneous authors to investigate the spread of this shift. In terms of more specific stylistic differences, we were able to find some common trends in both James and Twain, not found in the reference language, such as a decrease in the use of body references and a very marked difference for plural cases. This could suggest that James and Twain focus much more on the individual than was common for their time, but also that this particularity decreased over time. Existential constructions seemed to have generally gained more popularity over time in all three sets, with this being particularly pronounced in James and Twain.

Our analysis of Henry James and Mark Twain with a focus on stylistic changes has highlighted a number of differences between them, as for instance their use of female pronouns. James seems to have been highly progressive in his focus on the female perspective. This view is also supported by Baym (1981), who believes that James posed a continual challenge to the masculinist bias of American critical theory. An interesting aspect to consider as part of this investigation would be to compare James' style to a British reference corpus, given that he spent the latter part of his life in Europe.

In terms of syntactic style, there are a number of differences, one of which being that James seems to compose much more intricate sentences than Twain, especially towards the end of his life, as has

already been identified by literary scholars. In general, Twain's language is more pessimistic, questioning, and contains many more religious references than James' texts. From a more topic-based point of view, one might also consider frequent themes discussed as part of their works and possible changes in them over time. Overall, what one might say about Twain and James is that although they appear to often use the same tools, they apply them very differently. Regarding general differences from the reference corpus, it is probable that James and Twain did not really conform to the language of their time, although this would need to be verified by looking at the works of authors with comparable lifespans.

8 CONCLUSION

This work considered salient features of language change in the works of two prominent American authors, Henry James and Mark Twain, as well as in a reference corpus. We were able to identify a number of interesting changes in both lexical and syntactic features, suggesting other possible leads to explore. As style is a very general concept encompassing a multitude of possible dimensions, we were only able to 'scratch the surface', and more experiments should follow, to continue this work. The earlier part of this paper outlines only one method of discovery for salient features, but others should be considered and investigated. This work highlights the importance of using a reference corpus to verify that any change perceived in an author's style is indeed only to be found in the work of that author.

ACKNOWLEDGEMENT

We would like to thank our anonymous reviewers for their helpful suggestions on how to improve the earlier version of this paper. Further, we would also like to thank Carmela Chateau Smith, whose thorough work has greatly improved this paper's readability and consistency. This research is supported by Science Foundation Ireland (SFI) through the CNGL Programme (Grants 12/CE/I2267 and 13/RC/2106) in the ADAPT Centre (`www.adaptcentre.ie`).

REFERENCES

Alex AYRES (2010), *The Wit and Wisdom of Mark Twain*, Harper Collins.

Nina BAYM (1981), Melodramas of Beset Manhood: How Theories of American Fiction Exclude Women Authors, *American Quarterly*, 33(2):123–139, ISSN 00030678, 10806490, http://www.jstor.org/stable/2712312.

Joseph Warren BEACH (1918), *The Method of Henry James*, Yale University Press.

Walter BLAIR (1963), Reviewed Work: Twain and the Image of History by Roger B. Salomon, *American Literature*, 34(4):578–580, http://www.jstor.org/stable/2923090.

Van Wyck BROOKS (1920), The Ordeal of Mark Twain, *New York: Dutton*.

Henry Seidel CANBY (1951), *Turn West, Turn East: Mark Twain and Henry James*, Biblo & Tannen Publishers.

Walter DAELEMANS (2013), Explanation in Computational Stylometry, in *Computational Linguistics and Intelligent Text Processing*, pp. 451–462, Springer.

Mark DAVIES (2012), The 400 Million Word Corpus of Historical American English (1810–2009), in *English Historical Linguistics 2010: Selected Papers from the Sixteenth International Conference on English Historical linguistics (ICeHl 16), Pécs, 23–27 August 2010*, pp. 231–61.

Maciej EDER, Mike KESTEMONT, and Jan RYBICKI (2013), Stylometry with R: A Suite of Tools, in *Digital Humanities 2013: Conference Abstracts*, pp. 487–89, University of Nebraska–Lincoln, Lincoln, NE, http://dh2013.unl.edu/abstracts/.

Jerome FRIEDMAN, Trevor HASTIE, and Robert TIBSHIRANI (2001), *The Elements of Statistical Learning*, volume 1, Springer Series in Statistics Springer, Berlin.

Jerome FRIEDMAN, Trevor HASTIE, and Robert TIBSHIRANI (2010), Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, 33(1):1–22, http://www.jstatsoft.org/v33/i01/.

David L HOOVER (2007), Corpus Stylistics, Stylometry, and the Styles of Henry James, *Style*, 41(2):174–203.

Gareth JAMES, Daniela WITTEN, Trevor HASTIE, and Robert TIBSHIRANI (2013), *An Introduction to Statistical Learning*, volume 112, Springer.

Henry JAMES (1884), *The Art of Fiction*, Longmans, Green and Company.

Timothy P. JURKA, Loren COLLINGWOOD, Amber E. BOYDSTUN, Emiliano GROSSMAN, and Wouter VAN ATTEVELDT (2012), *RTextTools: Automatic Text Classification via Supervised Learning*, http://CRAN.R-project.org/package=RTextTools, R package version 1.3.9.

Michael J. Kane, John Emerson, and Stephen Weston (2013), Scalable Strategies for Computing with Massive Data, *Journal of Statistical Software*, 55(14):1–19, http://www.jstatsoft.org/v55/i14/.

Carmen Klaussner and Carl Vogel (2015), Stylochronometry: Timeline Prediction in Stylometric Analysis, in Max Bramer and Miltos Petridis, editors, *Research and Development in Intelligent Systems XXXII*, pp. 91–106, Springer International Publishing, Cham.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon (2011), Authorship Attribution in the Wild, *Language Resource Evaluation*, 45(1):83–94, doi:10.1007/s10579-009-9111-2, http://dx.doi.org/10.1007/s10579-009-9111-2.

Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow (2007), Measuring Differentiability: Unmasking Pseudonymous Authors, *Journal of Machine Learning Resources*, 8:1261–1276, ISSN 1532-4435, http://dl.acm.org/citation.cfm?id=1314498.1314541.

Max Kuhn (2014), *Caret: Classification and Regression Training*, http://CRAN.R-project.org/package=caret, with contributions from: Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer and the R Core Team, R package version 6.0-30.

Spyros Makridakis, Steven C Wheelwright, and Rob J Hyndman (2008), *Forecasting Methods and Applications*, John Wiley & Sons.

Meik Michalke (2014), *koRpus: An R Package for Text Analysis*, http://reaktanz.de/?c=hacking&s=koRpus, (Version 0.05-4).

James W Pennebaker and Lori D Stone (2003), Words of Wisdom: Language Use Over the Life Span, *Journal of Personality and Social Psychology*, 85(2):291–231.

Revolution Analytics and Steve Weston (2014), *foreach: Foreach looping construct for R*, http://CRAN.R-project.org/package=foreach, R package version 1.4.2.

Paolo Rosso, Francisco M. Rangel Pardo, Martin Potthast, Efstathios Stamatatos, Michael Tschuggnall, and Benno Stein (2016), Overview of PAN'16 – New Challenges for Authorship Analysis: Cross-Genre Profiling, Clustering, Diarization, and Obfuscation, in *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5–8, 2016, Proceedings*, pp. 332–350.

Helmut Schmid (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, in *Proceedings of International Conference on New Methods in Language Processing*, volume 12, pp. 44–49, Manchester, UK.

Joseph A. SMITH and Colleen KELLY (2002), Stylistic Constancy and Change across Literary Corpora: Using Measures of Lexical Richness to Date Works, *Computers and the Humanities*, 36(4):411–430, http://www.jstor.org/stable/30204686.

Efstathios STAMATATOS (2012), On the Robustness of Authorship Attribution Based on Character N-gram Features, *Journal of Law & Policy*, 21:421–439.

THOMAS M. WALSH AND THOMAS D. ZLATIC (1981), Mark Twain and the Art of Memory, *American Literature*, 53(2):214–231, http://www.jstor.org/stable/2926100.

James D. WILLIAMS (1965), The Use of History in Mark Twain's 'A Connecticut Yankee', *PMLA*, 80(1):102–110, http://www.jstor.org/stable/461131.

Hui ZOU and Trevor HASTIE (2005), Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, http://www.jstor.org/stable/3647580.