



Journal of Language Modelling

VOLUME 8 ISSUE 2
DECEMBER 2020



*Institute of Computer Science
Polish Academy of Sciences
Warsaw*

Journal of Language Modelling

VOLUME 8 ISSUE 2
DECEMBER 2020

Articles

Constraint summation in phonological theory 251
Giorgio Magri, Benjamin Storme

Word prediction in computational historical linguistics 295
Peter Dekker, Willem Zuidema

Serial verb constructions and covert coordinations in Edo
– an analysis in Type Logical Grammar 337
Ralf Naumann, Thomas Gamerschlag

Tools and resources

A French corpus annotated for
multiword expressions and named entities 415
*Marie Candito, Mathieu Constant, Carlos Ramisch,
Agata Savary, Bruno Guillaume, Yannick Parmentier,
Silvio Ricardo Cordeiro*



JOURNAL OF
LANGUAGE MODELLING

ISSN 2299-8470 (electronic version)

ISSN 2299-856X (printed version)

<http://jlm.ipipan.waw.pl/>

MANAGING EDITOR

Adam Przepiórkowski IPI PAN

SECTION EDITORS

Elżbieta Hajnicz IPI PAN

Agnieszka Mykowiecka IPI PAN

Marcin Woliński IPI PAN

STATISTICS EDITOR

Łukasz Dębowski IPI PAN



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 X_YL^AT_EX using the typefaces: *Playfair*
by Claus Eggers Sørensen, *Charis SIL* by SIL International,
JLM monogram by Łukasz Dziedzic.

*All content is licensed under
the Creative Commons Attribution 4.0 International License.*

 <http://creativecommons.org/licenses/by/4.0/>

EDITORIAL BOARD

Steven Abney University of Michigan, USA

Ash Asudeh Carleton University, CANADA;
University of Oxford, UNITED KINGDOM

Chris Biemann Technische Universität Darmstadt, GERMANY

Igor Boguslavsky Technical University of Madrid, SPAIN;
Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, RUSSIA

António Branco University of Lisbon, PORTUGAL

David Chiang University of Southern California, Los Angeles, USA

Greville Corbett University of Surrey, UNITED KINGDOM

Dan Cristea University of Iași, ROMANIA

Jan Daciuk Gdańsk University of Technology, POLAND

Mary Dalrymple University of Oxford, UNITED KINGDOM

Darja Fišer University of Ljubljana, SLOVENIA

Anette Frank Universität Heidelberg, GERMANY

Claire Gardent CNRS/LORIA, Nancy, FRANCE

Jonathan Ginzburg Université Paris-Diderot, FRANCE

Stefan Th. Gries University of California, Santa Barbara, USA

Heiki-Jaan Kaalep University of Tartu, ESTONIA

Laura Kallmeyer Heinrich-Heine-Universität Düsseldorf, GERMANY

Jong-Bok Kim Kyung Hee University, Seoul, KOREA

Kimmo Koskenniemi University of Helsinki, FINLAND

Jonas Kuhn Universität Stuttgart, GERMANY

Alessandro Lenci University of Pisa, ITALY

Ján Mačutek Comenius University in Bratislava, SLOVAKIA

Igor Mel'čuk University of Montreal, CANADA

Glyn Morrill Technical University of Catalonia, Barcelona, SPAIN

Stefan Müller Freie Universität Berlin, GERMANY

Mark-Jan Nederhof University of St Andrews, UNITED KINGDOM

Petya Osenova Sofia University, BULGARIA

David Pesetsky Massachusetts Institute of Technology, USA

Maciej Piasecki Wrocław University of Technology, POLAND

Christopher Potts Stanford University, USA

Louisa Sadler University of Essex, UNITED KINGDOM

Agata Savary Université François Rabelais Tours, FRANCE

Sabine Schulte im Walde Universität Stuttgart, GERMANY

Stuart M. Shieber Harvard University, USA

Mark Steedman University of Edinburgh, UNITED KINGDOM

Stan Szpakowicz School of Electrical Engineering
and Computer Science, University of Ottawa, CANADA

Shravan Vasishth Universität Potsdam, GERMANY

Zygmunt Vetulani Adam Mickiewicz University, Poznań, POLAND

Aline Villavicencio Federal University of Rio Grande do Sul,
Porto Alegre, BRAZIL

Veronika Vincze University of Szeged, HUNGARY

Yorick Wilks Florida Institute of Human and Machine Cognition, USA

Shuly Wintner University of Haifa, ISRAEL

Zdeněk Žabokrtský Charles University in Prague, CZECH REPUBLIC

Constraint summation in phonological theory*

Giorgio Magri¹ and Benjamin Storme²

¹ CNRS, University of Paris 8

² University of Lausanne

ABSTRACT

The classical constraints used in phonological theory apply to a single candidate at a time. Yet, some proposals in the phonological literature have enriched the classical constraint toolkit with constraints that instead apply to multiple candidates simultaneously. For instance, Dispersion Theory (Flemming 2002, 2004, 2008) adopts *distinctiveness constraints* that penalize pairs of surface forms which are not sufficiently dispersed. Also, some approaches to paradigm uniformity effects (Kenstowicz 1997; McCarthy 2005) adopt *Optimal Paradigm faithfulness constraints* that penalize pairs of stems in a paradigm which are not sufficiently similar. As a consequence, these approaches need to “lift” the classical constraints from a single candidate to multiple candidates by summing constraint violations across multiple candidates. Is this assumption of constraint summation typologically innocuous? Or do the classical constraints make different typological predictions when they are summed, independently of the presence of distinctiveness or optimal paradigm faithfulness constraints? The answer depends on the underlying model of constraint optimization, namely on

Keywords:
constraint-based
phonology,
Optimality
Theory,
Harmonic
Grammar,
Dispersion Theory,
paradigm
uniformity,
Optimal
Paradigms model

*We would like to thank Alan Prince for very useful feedback. We also would like to thank Edward Flemming for calling our attention to McCarthy (2005) as another application of constraint summation. The research reported in this paper has been supported by a grant from the Agence National de la Recherche (project title: The mathematics of segmental phonotactics) and by a grant from the MIT France Seed Fund (project title: Phonological Typology and Learnability).

how the profiles of constraint violations are ordered to determine the smallest one. Extending an independent result by Prince (2015), this paper characterizes those orderings for which the assumption of constraint summation is typologically innocuous. As a corollary, the typological innocuousness of constraint summation is established within both Optimality Theory and Harmonic Grammar.

1

INTRODUCTION

The classical constraints used in the phonological literature evaluate individual candidate surface realizations of a given underlying form (Prince and Smolensky 1993/2004). Yet, some authors have extended this classical constraint toolkit through constraints that evaluate not a *single* candidate but *multiple* candidates simultaneously. One example is provided by *distinctiveness* constraints in *Dispersion Theory* (DT; Flemming 2002, 2004, 2008), which penalize surface forms which are not sufficiently contrastive. Another example is provided by *Optimal Paradigm* (OP) faithfulness constraints in theories of paradigm uniformity effects such as the *Optimal Paradigms model* (OPM; Kenstowicz 1997; McCarthy 2005), which penalize dissimilarities among surface forms in an inflectional paradigm.

The addition of distinctiveness and OP faithfulness constraints to the classical constraint set raises a subtle technical problem: since classical constraints apply to a single candidate at a time while distinctiveness and OP faithfulness constraints instead apply to multiple candidates simultaneously, the classical constraints need to be “lifted” from individual candidates to tuples of candidates. A natural solution to this problem (and indeed the solution pursued in DT and the OPM) is to lift a classical constraint *C* from individual candidates to tuples of candidates by summing constraint violations across the candidates in the tuple, as in (1).¹

¹ In order for this assumption (1) to make sense, the sum on the right-hand side must be finite. Finiteness requires one of two conditions to hold: either the sum has only a finite number of addenda; or else it has an infinite number of addenda but only finitely many of them are different from zero because the con-

(1)

Constraint summation assumption:

$$C(\langle \text{candidate 1, candidate 2, candidate 3} \dots \rangle) = \\ = C(\text{candidate 1}) + C(\text{candidate 2}) + C(\text{candidate 3}) + \dots$$

To set the stage for the paper, Section 2 reviews the arguments for this constraint summation assumption (1) in DT and the OPM.

The use of distinctiveness constraints to model contrast is a topic of intense debate in the current phonological literature (see for instance Blevins 2004; Boersma and Hamann 2008; Stanton 2017), as is the use of OP faithfulness constraints to capture paradigm uniformity effects (see for instance Albright 2010). This paper contributes to these debates by taking a closer look at a formal consequence of these constraints, namely the assumption (1) that classical constraints be lifted through constraint summation. What are the phonological implications of this assumption? To zoom in on this question, let us suppose that the constraint set contains no distinctiveness or OP faithfulness constraints but only classical constraints. We then have two options. According to the classical approach, we can compute the optimal surface realization of each underlying form individually relative to the original classical constraints. Alternatively, we can compute the optimal surface realizations for all the underlying forms simultaneously

straint C only penalizes finitely many of the candidates considered. As we will see in Section 2, in the case of the OPM, the number of addenda on the right-hand side of (1) is indeed finite because it is controlled by the size of the inflectional paradigm, which is a finite set of forms. For applications of DT to segment inventories, the number of addenda on the right-hand side of (1) is also plausibly finite, because it is controlled by the size of the underlying universal inventory of atomic linguistic sounds, which is plausibly finite. Yet, for applications of DT to strings of segments, the number of addenda on the right-hand side of (1) is controlled by the number of strings, which is infinite unless we can cap their length in some phonologically plausible way. Furthermore, it is unlikely that a constraint would penalize only finitely many candidates in this case, as pointed out to us by Edward Flemming (p.c.). For example, if C is a markedness constraint penalizing voiced stops, it will assign violations to the infinite set of strings containing voiced stops. Similarly, if C is an identity faithfulness constraint for voicing, it will be violated by an infinite number of mappings with voiced stops in the input string corresponding to voiceless stops in the output string. We leave this technical issue open at this stage.

relative to the summed version (1) of the classical constraints. Is it the case that these two approaches lead to the same set of winners, so that constraint summation is innocuous? Equivalently, is it the case that phonological theories that make use of constraint summation, such as DT and the OPM, actually coincide with classical constraint-based phonology when the constraint set consists solely of classical constraints but no distinctiveness or OP faithfulness constraints? Section 3 formalizes this question.

Obviously, the individual constraint violations $C(\textit{candidate 1})$ and $C(\textit{candidate 2})$ cannot be reconstructed from their sum $C(\textit{candidate 1}) + C(\textit{candidate 2})$. One might thus intuitively expect that the assumption (1) of constraint summation wipes away much of the information encoded by the classical constraints. If that were indeed the case, phonological frameworks such as DT and the OPM which make use of constraint summation could profoundly alter the typological implications of the classical constraints, possibly leading to pernicious typological predictions. The goal of this paper is to show that this pessimism is unwarranted.

To start, Section 4 focuses on the case of *Optimality Theory* (OT; Prince and Smolensky 1993/2004). In OT, violation profiles are optimized relative to the *lexicographic order* induced by some constraint *ranking*. In the context of OT, the typological innocuousness of the assumption (1) of constraint summation has been established in Prince (2015). Interestingly, we observe that Prince's original proof can be substantially simplified if we reason in terms of violation profiles rather than in terms of *elementary ranking conditions* (ERCs; Prince 2002), as Prince does. The fact that an OT-specific tool like ERCs hinders rather than facilitates the proof suggests that Prince's result must be independent of the specifics of OT and instead follow from some deeper, more general structure. What is this structure?

The statistician Michel Talagrand explains why it is important to pursue this question: "The practitioner [...] is likely to be struggling at any given time with his favorite model of the moment, a model that will typically involve a rather rich and complicated structure. There is a near infinite supply of such models. Fashions come and go, and the importance with which we view any specific model is likely to strongly vary over time. [One should thus] always consider a problem under the minimum structure in which it makes sense. [...]"

By following [this advice], one is naturally led to the study of problems with a kind of minimal and intrinsic structure. Besides the fact that it is much easier to find the crux of the matter in a simple structure than in a complicated one, there are not so many really basic structures, so one can hope that they will remain of interest for a very long time.” (Talagrand 2014)

Pursuing this insight, Section 5 offers a complete (both necessary and sufficient) characterization of the “minimal structure” needed to guarantee the typological innocuousness of the assumption (1) of constraint summation, namely the structure provided by additive weak orders. This characterization shows that OT’s specific choice of the lexicographic order is by no means necessary to ensure the typological innocuousness of constraint summation. As discussed in Section 6, typological innocuousness indeed extends beyond OT to a variety of constraint-based optimization schemes, crucially including optimization schemes based on additive utility functions, as in *Linear OT* (LOT; Keller 2000, 2006) and *Harmonic Grammar* (HG; Legendre *et al.* 1990b,a; Smolensky and Legendre 2006), which have figured prominently in the recent phonological literature (Pater 2009; Potts *et al.* 2010). Section 7 concludes the paper by discussing the implications of this result for the formal foundations of phonological approaches that make use of constraint summation, such as DT and the OPM.

WHY IS CONSTRAINT SUMMATION NEEDED IN PHONOLOGICAL THEORY

2

To set the stage for the paper, this section reviews arguments for the constraint summation assumption (1) in DT and in the OPM.² Our presentation stresses the complete formal analogy between the two arguments, despite the fact that they belong to distant corners of phonological theory. The rest of the paper will then take a closer look at the constraint summation assumption (1) motivated in this section.

²The reader already familiar with DT and the OPM might want to skip ahead to Section 3.

2.1

Dispersion Theory

This section summarizes the argument for constraint summation in DT. The argument has three steps. First, Subsection 2.1.1 reviews Flemming’s challenge against classical markedness and faithfulness constraints that look at a single candidate at a time. Second, Subsection 2.1.2 overviews Flemming’s proposal that the classical constraint toolkit be enriched with *distinctiveness* constraints that look at multiple candidates simultaneously. Third, Subsection 2.1.3 illustrates how the classical constraints are “lifted” to multiple candidates through constraint summation (1) in order for them to be able to interact with distinctiveness constraints.³

2.1.1

Insufficiency of classical
markedness and faithfulness constraints

The constraint-based phonological literature assumes two classes of constraints. *Faithfulness* constraints measure the distance or discrepancy between an underlying form and its surface realization. *Markedness* constraints measure the phonotactic ill-formedness of a surface form. Both types of constraints thus look at a single underlying/surface form candidate pair at a time. Flemming (2002, 2004) argues that this classical toolkit is insufficient. We review here one of his arguments, based on the typology of systems of contrasts among voiceless, plain voiced, and prenasalized voiced stops (Flemming 2004, pages 258–263). Many languages contrast voiceless stops [p, t, k] with plain voiced stops [b, d, g] (e.g. French; Tranel 1987). Yet there are also a few languages that prefer having prenasalized voiced stops [ⁿb, ⁿd, ⁿg]

³The architecture summarized in Subsections 2.1.2–2.1.3 is a simplified version of the architecture proposed in Flemming (2008) (not the earlier one proposed in Flemming 2002). Our presentation is simplified because it confounds Flemming’s (2008) three modules into a single one. In fact, we ignore Flemming’s orthogonal claim that language specific properties of phonetic realization play a role in the phonology. Hence, we conflate Flemming’s *realized inputs* with underlying phonological forms and effectively ignore the “phonetic realization module” which derives the former from the latter. Furthermore, we ignore the distinction between the “inventory selection module” and the “phonotactics module”, following Flemming (2017b,a). These simplifications are adopted for simplicity only and they do not affect the scope of our results.

(instead of plain voiced stops) contrast with voiceless stops (e.g. San Juan Colorado Mixtec; Iverson and Salmons 1996). How could such a language be derived with classical markedness and faithfulness constraints?

Obviously, we would need a markedness constraint which penalizes plain voiced stops at the exclusion of prenasalized ones. Let us call this constraint *D. The intuition behind this constraint could be that voicing is harder to sustain in a plain voiced stop than in a prenasalized one because the nasal aperture facilitates voicing by preventing a fast pressure buildup above the glottis (Ohala 1983). We assume that this constraint *D is “counterbalanced” by another markedness constraint *ⁿD that instead penalizes prenasalized voiced stops at the exclusion of plain ones. The intuition behind this constraint would be that prenasalized stops are more effortful to produce because they “require rapid raising of the velum to produce oral and nasal phases within the same stop” (Flemming 2004, page 260). Finally, we consider a third markedness constraint *VTV which penalizes voiceless stops in intervocalic position. The constraint set is completed by two faithfulness constraints Ident(voice) and Ident(nas) that protect the underlying specifications for voicing and nasalization, respectively.

For concreteness, let us adopt the OT model of constraint interaction (see Section 4 below for a review of the OT formalism). Tableau (2a) derives the faithful realization of underlying voiceless stops intervocalically. And tableau (2b) derives prenasalization of an underlying plain voiced stop.

(2) a.

/ata/	*D	Ident(voice)	Ident(nas)	* ⁿ D	*VTV
☞ [ata]					*
[ada]	*!	*			
[a ⁿ da]		*!	*	*	

b.

/ada/	*D	Ident(voice)	Ident(nas)	* ⁿ D	*VTV
[ata]		*!			*
[ada]	*!				
☞ [a ⁿ da]			*	*	

In conclusion, we have derived a language like San Juan Colorado Mixtec, where voiceless stops contrast with prenasalized voiced stops but not with plain voiced stops.

However, Flemming notes that no language prefers prenasalized voiced stops to plain voiced stops in a context where voiceless stops are banned. For instance, intervocalic voiceless stops are never repaired through intervocalic prenasalization. The only attested repair is intervocalic voicing (e.g. Tümpisa Shoshone; Dayley 1989). Flemming argues that this fact is difficult to derive with only the faithfulness and markedness constraints made available by classical constraint-based phonology. In fact, as soon as *D is allowed to outrank *ⁿD as in tableaux (2), we derive an unattested pattern of intervocalic prenasalization of voiceless stops. This pattern is derived if *VTV and Ident(voice) are flipped, as in (3).

(3)

/ata/	*VTV	*D	* ⁿ D	Ident(voice)	Ident(nas)
[ata]	*!				
[ada]		*!		*	
ᵀᵀ [a ⁿ da]			*	*	*

These considerations suggest that our initial attempt at deriving the preference for prenasalized over plain voiced stops in Mixtec through classical markedness constraints is not on the right track. The constraint responsible for this preference cannot be a classical markedness constraint such as the constraint *D proposed above, because that constraint is blind to the presence or absence of a plain voiceless stop. This strategy based on classical markedness thus leads to the incorrect prediction that prenasalized voiced stops are preferred also in the absence of plain voiceless stops, namely that prenasalization can be used as a repair strategy for intervocalic voiceless stops.

2.1.2

Distinctiveness constraints

In order to solve this impasse, Flemming proposes that the preference for prenasalized voiced stops in contexts where voiceless stops are available results from contrast enhancement: the voicing contrast is more distinct in the pair [t]-[ⁿd] than in the pair [t]-[d] (Iverson and Salmons 1996), due to the higher intensity of the periodic part of the speech signal in [ⁿd] than in [d]. In the presence of a voiceless stop,

the preference for maximizing contrast can exert its effect and allow for [ʰd] at the exclusion of plain [d]. But in the absence of voiceless stops, there is no contrast to enhance and thus the markedness of [ʰd] relative to [d] is the only active force, whereby voiced stops are predicted to be systematically preferred.

Flemming formalizes the preference for more distinct contrasts via *distinctiveness* constraints that penalize pairs of sounds based on their distance along a perceptual scale. In the case at hand, the relevant perceptual scale is the intensity of voicing. Following Flemming's simplifying assumption, suppose that the intensity of voicing is equal to 0 in voiceless stops, to 1 in plain voiced stops, and to 2 in prenasalized stops. Pairs [t]-[d] and [d]-[ʰd] (but not [t]-[ʰd]) violate a distinctiveness constraint requiring voicing contrasts corresponding to a distance strictly larger than one unit along the intensity scale. This constraint is denoted MinDist, as in (4).

(4) MinDist:

Assign a violation mark to pairs of surface forms with a voicing contrast corresponding to a distance equal to or smaller than 1 along the scale of voicing intensity.

Penalizes [t]-[d], [d]-[ʰd]. Does not penalize [t]-[ʰd].

All three pairs [t]-[ʰd], [t]-[d], and [d]-[ʰd] violate a distinctiveness constraint requiring voicing contrasts corresponding to a distance strictly larger than two units along the intensity scale (we ignore this constraint in what follows because it does not distinguish among these three pairs).

Lifting classical constraints through constraint summation

2.1.3

Distinctiveness constraints are formally very different from classical faithfulness and markedness constraints. In fact, classical constraints assign a number of violations to each individual candidate surface realization of a given underlying form. Distinctiveness constraints instead compare tuples of surface realizations of multiple underlying forms. This difference has implications for the architecture of grammar. A classical grammar in the constraint-based literature evaluates the candidates of a single underlying form at a time, as illustrated above with the two separate tableaux (2) for the two underlying forms /ata/ and /ada/. A grammar with distinctiveness constraints instead

must evaluate tuples of candidates corresponding to multiple underlying forms. But what about the classical constraints that are now mixed up with the distinctiveness constraints? How can they be “lifted” from individual candidates to tuples of candidates of multiple underlying forms? Flemming makes the natural suggestion that classical faithfulness and markedness constraints be redefined for tuples of candidates by summing their constraint violations across all candidates in the tuple, as anticipated in (1).

Tableau (5) illustrates how distinctiveness constraints and constraint summation of the classical constraints work in DT. We consider again the two underlying forms /ata/ and /ada/. This time, they occur together in the same tableau, rather than heading the two separate tableaux in (2). These two underlying forms have three surface candidates [ata], [ada], and [aⁿda] each in the classical approach of tableaux (2). In DT, we thus consider 3 × 3 = 9 pairs of candidates, listed by row in (5). For instance, row (5d) corresponds to the (impossible) mapping whereby /ata/ is realized as [ada] and /ada/ as [ata].

(5)

/ata/, /ada/	MinDist	Ident(voice)	Ident(mas)	* _{TD}	* _D	* _{VTV}
a. [ata], [ata]		* _{d→t}				* _{ata} * _{ata}
b. [ata], [ada]	* _{t-d}				* _d	* _{ata}
c. [ata], [a ⁿ da]			* _{d→nd}	* _{nd}	* _d	* _{ata}
d. [ada], [ata]	* _{t-d}	* _{t→d} * _{d→t}			* _d	* _{ata}
e. [ada], [ada]		* _{t→d}			* _d * _d	
f. [ada], [a ⁿ da]	* _{d-nd}	* _{t→d}	* _{d→nd}	* _{nd}	* _d * _{nd}	
g. [a ⁿ da], [ata]		* _{t→nd} * _{d→t}	* _{t→nd}	* _{nd}	* _d	* _{ata}
h. [a ⁿ da], [ada]	* _{d-nd}	* _{t→nd}	* _{t→nd}	* _{nd}	* _d * _{nd}	
i. [a ⁿ da], [a ⁿ da]		* _{t→nd}	* _{t→nd} * _{d→nd}	* _{nd} * _{nd}	* _{nd} * _{nd}	

The distinctiveness constraint MinDist penalizes the pair of surface forms in (5b), because their consonants sit on the voicing scale at a distance of 1 ([ata]-[ada]), respectively. It does not penalize the pair

of surface forms in (5c), because their consonants sit on the voicing scale sufficiently far apart, namely at a distance of 2 ([ata]-[aⁿda]). And so on. This constraint thus exerts a preference for the prenasalized over the plain voiced stop, although crucially only in the presence of the voiceless stop.

Classical markedness and faithfulness constraints are summed across multiple candidates. For instance, a classical faithfulness constraint such as Ident(voice) assigns two violations to the pair of surface forms in (5d), because it assigns one violation to the mapping of /ata/ to [ada] and another violation to the mapping of /ada/ to [ata] and the two violations are summed together, as prescribed by the constraint summation assumption (1). As another example, a classical markedness constraint such as *VTV assigns two violations to the pair of surface forms in (5a), because it features two instances of the surface form [ata]. And so on. To make it easier to track constraint violations, the specific pairs of output segments (in the case of distinctiveness constraints), single output segments (in the case of classical markedness constraints), and input-output segments (in the case of classical faithfulness constraints) that violate the corresponding constraint are indicated in subscript next to each violation mark.⁴

This approach solves the problem discussed in Subsection 2.1.1: it derives a system contrasting voiceless and prenasalized voiced stops while blocking allophonic prenasalization of voiceless stops. In fact, a system with contrasting voiceless and prenasalized voiced stops is derived if MinDist and Ident(voice) are top ranked: this ranking condition eliminates all options but for the desired option ⟨[ata], [aⁿda]⟩ in row (5c). Furthermore, nasalization as a repair to intervocalic voiceless stops is impossible because the three logically possible options that prenasalize intervocalic voiceless /t/ are all harmonically bounded. In fact, the option ⟨[aⁿda], [ata]⟩ in row (5g) is

⁴As anticipated in the informal discussion at the beginning of Subsection 2.1.2, Flemming assumes that the MinDist distinctiveness constraint is the only constraint that favors prenasalized over plain voiced stops, while classical markedness constraints prefer plain over prenasalized voiced stops. The markedness constraint *D thus needs to be redefined as penalizing all voiced stops, both plain and prenasalized ones. It therefore assigns two violations in (5f), because its two surface forms [ada] and [aⁿda] both violate it.

harmonically bounded by ⟨[ata], [aⁿda]⟩ in row (5c). And the options ⟨[aⁿda], [ada]⟩ and ⟨[aⁿda], [aⁿda]⟩ in rows (5h) and (5i) are both harmonically bounded by ⟨[ada], [ada]⟩ in row (5e).

2.2 Optimal Paradigms model

This section summarizes the argument for constraint summation in the OPM. The argument has three steps, in complete analogy with the preceding Subsection 2.1. First, Subsection 2.2.1 reviews McCarthy’s 2005 challenge that some inflectional paradigms raise for asymmetric, base-prioritizing theories of output-output correspondence. Second, Subsection 2.2.2 overviews McCarthy’s proposal that the classical constraint toolkit be enriched with OP faithfulness constraints that evaluate all paradigm members simultaneously. Third, Subsection 2.2.3 illustrates how the classical constraints are “lifted” to entire paradigms through constraint summation (1) in order for them to be able to interact with OP faithfulness constraints.

2.2.1 Insufficiency of asymmetric output-output faithfulness constraints

Morphologically-related forms may bear resemblance that goes beyond what is predicted by the interaction of classical markedness constraints and input-output faithfulness constraints. A classical example is the case of the participle *lightening* [laɪtɪŋ], where the stem-final consonant is realized as a syllabic nasal [ŋ], as in the verb *lighten* [laɪtɪ], instead of the phonotactically expected [n]. Data of this kind have motivated positing another type of faithfulness besides input-output faithfulness: output-output faithfulness.⁵ Output-output faithfulness constraints enforce similarity among surface forms. In the case of surface inflected forms, similarity is enforced among surface forms in the same inflectional paradigm, i.e. forms that share a lexeme. In this approach, the presence of syllabic [ŋ] in the participle *lightening* [laɪtɪŋ] can be explained as the result of an output-output faithfulness constraint requiring similarity with the verb *lighten* [laɪtɪ] and

⁵Output-output faithfulness is also motivated by patterns of reduplication (McCarthy and Prince 1995).

outranking the input-output faithfulness constraint requiring similarity with the input stem /lartn/.

When one form in a paradigm is morphologically simpler than other paradigm members, this morphologically simpler form (i.e. the base) is always the one that other paradigm members must be faithful to (see Benua 1997, 240–242 for a discussion of potential counterexamples). *Lightening* conforms to this generalization because it is asymmetrically influenced by its base *lighten*. In line with this generalization, theories of output-output faithfulness have been developed where the phonology of the base is computed in a first step and serves as input to the evaluation of affixed forms, alongside the affixed forms' underlying representations (e.g. Benua's 1997 *Transderivational Correspondence Theory*).

However, McCarthy (2005) notes that effects that can be analyzed as output-output faithfulness are also observed in paradigms where all forms are equally complex morphologically and where the choice of the attractor is not guided by morphological simplicity or markedness but by phonological markedness. McCarthy (2005) illustrates his argument with Arabic verbal stems. In Arabic, verbal stems are required to end in VC (e.g. [faʕal], [faʕʕal]). No stem ending in V:C or VCC is attested in verbal paradigms (e.g. *[faʕa:l], *[faʕl]). This contrasts with nominal stems, which can end in VC, V:C, and VCC (e.g. [faʕal], [faʕa:l], [faʕl]). Under Richness of the Base, the fact that the phonological shape of verbal stems is more constrained than that of nominal stems is unexpected.

McCarthy's insight is that this apparent quirk of Arabic verbs can be explained as an effect of output-output faithfulness, combined with an independent property that distinguishes nouns and verbs in Arabic. Nominal suffixes in Arabic all start with a vowel whereas verbal suffixes start with vowels or consonants (see McCarthy 2005, 179-180 for a list of suffixes). In a nutshell, due to a high ranking markedness constraint banning super heavy syllables (*V:CCV, *VCCCV), verbal stems followed by consonant-initial suffixes can only afford short vowels in stem-final syllables (e.g. [faʕal-tu] but *[faʕa:l-tu] and *[faʕl-tu]). Output-output faithfulness then extends the short vowel that is phonotactically expected before consonant-initial suffixes to the whole paradigm, including to forms built with vowel-initial suffixes and where short vowels are not phonotactically required. In nouns,

only vowel-initial suffixes are attested. Therefore, contrary to inflected verbs, there is no paradigmatic pressure to extend stems ending in VC-, therefore allowing for all VC-, V:C-, and VCC- to surface faithfully in inflected nouns.

For this analysis to be implemented using Benua's Transderivational Correspondence Theory, it is necessary to assume that the base in verbal paradigms is one of the forms built with a consonant-initial suffix. Faithfulness to the base then extends the short vowel that is phonotactically expected in this form to all other forms. Tableaux (6a) and (6b) show how this analysis works, focusing on two forms of the paradigm of hypothetical underlying /faʃa:l/: (i) an inflected form with a consonant-initial suffix that serves as the base in the paradigm, /faʃa:l-tu/ (1st singular perfective), and (ii) an inflected form with a vowel-initial suffix, /faʃa:l-a/ (3d singular perfective).

(6) a.

/faʃa:l-tu/	*V:CCV	IdBD(length)	IdIO(length)
[faʃa:ltu]	*		
☞ [faʃaltu]			*

b.

/faʃa:l-a/ Base = [faʃaltu]	*V:CCV	IdBD(length)	IdIO(length)
[faʃa:la]		*	
☞ [faʃala]			*

To get the short vowel to be extended to other forms, /faʃa:l-tu/ has to be considered as the base. As the base, its phonology is computed first. In this first cycle, only input-output correspondence is relevant. Because *V:CCV outranks the faithfulness constraint protecting underlying vowel length (IdentIO(length)), the stem long vowel is shortened before CC, as shown in tableau (6a). In a second step, the phonology of other paradigm members is computed. Now, output-output correspondence is relevant, with [faʃa:ltu] serving as the base for the surface form derived from underlying /faʃa:l-a/. IdentBD(length) requires the form under evaluation to match the base along vowel length. This constraint outranks IdentIO(length), therefore favoring base-derivative similarity over input-output similarity, as shown in tableau (6b).

As pointed out by McCarthy, the problem with this approach is that there is no independent, morphological motivation to treat a form with a consonant-initial suffix (/faʕa:l-tu/ in our case) as the base. Indeed, inflected forms with consonant-initial suffixes are neither simpler than the others paradigm members (all forms are inflected) nor morphosyntactically less marked. Indeed, morphosyntactic markedness predicts that the third person singular form should be the base. However, all third person singular forms in the verbal paradigm are built with vowel-initial suffixes (cf. /faʕa:l-a/ in our example).

Optimal Paradigms faithfulness constraints

2.2.2

To solve this issue, McCarthy proposes *Optimal Paradigm* faithfulness constraints. Surface inflected forms are related by output-output correspondence to all other inflected forms of the same stem. The stem of every paradigm member stands in correspondence with the stem of other members; OP faithfulness constraints enforce similarity among corresponding stems in a paradigm. The resulting system is distinct from Benua's Transderivational Correspondence Theory because the latter is asymmetrical (the base is generated "first", hence not modifiable) while the effects of OP faithfulness are symmetric: all members in a paradigm are evaluated simultaneously hence each of them can be modified.

In the case of Arabic, extension of the short vowel from stems built with consonant-initial suffixes is enforced by an OP faithfulness constraint that requires matching vowel length in all pairs of paradigm members, as defined in (7).

(7) Ident-OP(length)

In every paradigm, the stem of each paradigm member corresponds to the stem of every other paradigm member.

In each pair of correspondent stems S1-S2, assign a violation mark for each vowel in S1 that does not have the same length as the corresponding vowel in S2.

Penalizes paradigm <faʕa:l-ta, faʕal-u> and <faʕal-ta, faʕa:l-u>. Does not penalize paradigms <faʕal-ta, faʕal-u> and <faʕa:l-ta, faʕa:l-u>

This constraint evaluates surface resemblance *symmetrically* across inflectionally related forms, hence it does not stipulate that any

paradigm member should be a priori preferred over the others. In a concrete analysis, the choice of the attractor is determined by markedness, as will be shown in more detail in the next subsection.

2.2.3 Lifting classical constraints through constraint summation

OP faithfulness constraints are formally very different from classical faithfulness and markedness constraints. In fact, classical constraints assign a number of violations to each individual candidate surface realization of a given underlying form. OP faithfulness constraints instead compare the surface realizations of multiple underlying forms based on their similarity. Again as in the case of DT discussed in the preceding subsection, this difference means that classical constraints need to be “lifted” from individual candidates to whole paradigms in order to be able to interact with OP faithfulness constraints. McCarthy (2005, p.173) makes the natural suggestion that classical faithfulness and markedness constraints be redefined by summing their constraint violations across all forms in a paradigm, as anticipated in (1).

Tableau (8) illustrates how OP faithfulness constraints and constraint summation of the classical constraints work in the OPM. We consider again the two underlying forms /faʃa:l-a/ and /faʃa:l-tu/. This time, they occur together in the same tableau, rather than heading the two separate tableaux in (6). These two underlying forms have two surface candidates each in the classical approach of tableaux (6). In the OPM, we thus consider $2 \times 2 = 4$ pairs of candidates, listed by row in (8). For instance, row (8a) corresponds to the mapping whereby /faʃa:l-a/ is realized as [faʃa:l-a] and /faʃa:l-tu/ as [faʃa:l-tu].

(8)

	*V:CCV	IdentOP(length)	IdentIO(length)
/faʃa:l-a/, /faʃa:l-tu/			
a. [faʃa:l-a], [faʃa:l-tu]	*		
b. [faʃa:l-a], [faʃal-tu]		*	*
c. [faʃal-a], [faʃa:l-tu]	*	*	*
☞ d. [faʃal-a], [faʃal-tu]			**

Classical markedness and faithfulness constraints are summed across multiple candidates. For instance, a classical faithfulness constraint such as IdentIO(length) assigns two violations to the pair of surface forms in row (8d), because it assigns one violation to the mapping of /faʕa:l-a/ to [faʕa:l-a] and another violation to the mapping of /faʕa:l-tu/ to [faʕa:l-tu] and the two violations are summed together, as prescribed by the constraint summation assumption (1).

The OP faithfulness constraint IdentOP(length) penalizes the pairs of surface forms in (8b) and (8c), because they feature two vowels that stand in correspondence but do not match in length. This constraint thus exerts a preference for paradigm uniformity, without specifying which form will be the attractor: the pairs of surface forms in (8a) and (8d) are equally good in terms of paradigm uniformity. The choice of the attractor is determined by the high ranked markedness constraint *V:CCV, which penalizes (8a) featuring a super heavy syllable. This approach solves the problem discussed in Subsection 2.2.1: it derives the generalization on verbal stems without needing to stipulate that inflected forms with consonant-initial suffixes are the base, in the morphological sense of Benua (1997). The reason why the short vowel length is extended to other forms rather than the long one is the fact that super heavy syllables are marked.

IS CONSTRAINT SUMMATION TYPOLOGICALLY INNOCUOUS?

3

The preceding section has reviewed some phonological theories (such as DT and the OPM) that share two formal innovations. The first innovation is that the classical constraint set is enriched with constraints (such as distinctiveness and OP faithfulness constraints) that evaluate multiple candidates *simultaneously* by comparing surface forms from the perspective of their distinctiveness or their mutual faithfulness. These constraints are therefore formally rather different from classical faithfulness and markedness constraints, which instead evaluate candidates *individually*, one at a time. The second related innovation is that classical markedness and faithfulness constraints are “lifted”

from an individual candidate to multiple candidates through the constraint summation assumption (1). Is this constraint summation assumption *typologically innocuous*? In other words, is it the case that constraint summation does not alter the typological implications of classical markedness and faithfulness constraints? Or is it instead the case that phonological theories (such as DT and the OPM) that make use of constraint summation predict very different typologies even when the constraint set consists only of classical constraints (but no distinctiveness or OP faithfulness constraints)? This section formulates this question explicitly.

3.1

The classical approach

Let us suppose that we have only two underlying forms. The reasoning developed in this and the following sections extends straightforwardly from two to an arbitrary finite number of underlying forms (see footnote 7 below; the extension to an infinite number of underlying forms is trickier, as discussed in footnote 1 above). In order to focus on the constraint summation assumption (1), we suppose that the constraint set consists of n classical constraints C_1, \dots, C_n , but no distinctiveness or OP faithfulness constraints. We denote by α the generic surface candidate of the first underlying form and we collect these surface candidates into a candidate set A . The classical constraints C_1, \dots, C_n assign to (the mapping of that underlying form into) the candidate α the n constraint violations a_1, \dots, a_n . We collect them together into a tuple $\mathbf{a} = (a_1, \dots, a_n)$. Analogously, we denote by β the generic surface candidate of the second underlying form and we collect these surface candidates into a candidate set B . The classical constraints C_1, \dots, C_n assign to (the mapping of that second underlying form into) the candidate β a tuple $\mathbf{b} = (b_1, \dots, b_n)$ of n constraint violations b_1, \dots, b_n . A concrete example of the sets A and B is provided by the two tableaux (2a) and (2b) for the two underlying forms /ata/ and /ada/. In this case, $n = 5$, the candidate corresponding to the first row of the tableau A is $\alpha = [\text{ata}]$, and the corresponding tuple of constraint violations is $\mathbf{a} = (0, 0, 0, 0, 1)$.

Under the assumption that the constraints suffice to capture all the relevant information, the optimal candidate for a given underlying form must be the one which violates the constraints the least, that

is which corresponds to the “smallest” tuple of constraint violations. To formalize this intuition, we extend the intuitive notion of “smaller than” from single numbers to tuples of numbers. Thus, the condition $\hat{a} < \mathbf{a}$ means that the tuple of constraint violations \hat{a} is smaller than the tuple of constraint violations \mathbf{a} . Effectively, this means that we define an *order* $<$ among tuples of constraint violations. Different implementations of (classical) constraint-based phonology considered in the literature differ for the choice of the order $<$ used to compare tuples of constraint violations. For full generality, we allow this order $<$ to be *partial*: for some pairs of tuples of constraint violations, $<$ might not be able to tell which one is smaller. In other words, some tuples might be *incommensurable*.⁶

We denote by $\text{opt}_{<} A$ the collection of *optimal* candidates in the set A , namely those candidates $\hat{\alpha}$ corresponding to a tuple $\hat{\mathbf{a}}$ of constraint violations which is minimal relative to the order $<$, as defined in (9).

- (9) $\text{opt}_{<} A$ is the set of those candidates $\hat{\alpha}$ in A such that there exists no competing candidate α in A such that the constraints assign to this competing candidate α a tuple \mathbf{a} of constraint violations which is smaller than the tuple $\hat{\mathbf{a}}$ of constraint violations they assign to the optimal candidate $\hat{\alpha}$, namely $\mathbf{a} < \hat{\mathbf{a}}$.

(Classical) constraint-based phonology assumes that the underlying form with candidate set A is mapped to a surface candidate $\hat{\alpha}$ which violates the constraints the least, namely which belongs to the optimal subset $\text{opt}_{<} A$ of candidates with the smallest tuples of constraint violations. Analogous considerations hold for the other underlying form, which is mapped to a surface candidate in the optimal set $\text{opt}_{<} B$.

We note that the set $\text{opt}_{<} A$ of optimal candidates can contain more than one candidate. In fact, two candidates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ can both be optimal if their corresponding tuples of constraint violations $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$ are incommensurable: neither of the two is larger than the other according to the order $<$ because $<$ is only partial. Furthermore, even

⁶ For instance, in the case of the HG implementation of constraint-based phonology (see Subsection 6.2 below), the order $<$ is defined in terms of a *utility* or *harmony function*. Two candidates with different tuples of constraint violations can achieve the same harmony. Their tuples of constraint violations are therefore incommensurable relative to this order $<$.

if $<$ is total, two different candidates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ can both be optimal if the constraints fail to distinguish between them, namely the two candidates share the same tuple $\hat{\mathbf{a}}_1 = \hat{\mathbf{a}}_2$ of constraint violations, as we will discuss in more detail in Subsection 3.4.

3.2 *The constraint summation approach of DT and the OPM*

We denote by $A \times B$ the collection of all pairs (α, β) of a candidate α from the candidate set A and a candidate β from the candidate set B . We “lift” the n classical constraints from single candidates to pairs of candidates through the constraint summation assumption (1). This means that the lifted constraints assign to the candidate pair (α, β) the component-wise sum $\mathbf{a} + \mathbf{b}$ of the tuples $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n)$ of constraint violations assigned to the two individual candidates α and β , as in (10).

$$(10) \quad \mathbf{a} + \mathbf{b} = (a_1 + b_1, \dots, a_k + b_k, \dots, a_n + b_n)$$

To illustrate, if A and B are the two tableaux (2a) and (2b), their product $A \times B$ is the tableau (11), which lists all pairs of candidates and sums the stars in the two corresponding cells of A and B .

(11)

	(/ata/, /ada/)	Ident(voice)	Ident(nas)	* ⁿ D	*D	*VTV
a.	([ata], [ata])	*				**
b.	([ata], [ada])				*	*
c.	([ata], [a ⁿ da])		*	*	*	*
d.	([ada], [ata])	**			*	*
e.	([ada], [ada])	*			**	
f.	([ada], [a ⁿ da])	*	*	*	**	
g.	([a ⁿ da], [ata])	**	*	*	*	*
h.	([a ⁿ da], [ada])	*	*	*	**	
i.	([a ⁿ da], [a ⁿ da])	*	**	**	**	

This is of course the same as tableau (5) considered above, stripped of the column corresponding to the distinctiveness constraint MinDist.

We denote by $\text{opt}_{<}(A \times B)$ the collection of *optimal* candidate pairs in the set $A \times B$, namely those candidate pairs $(\widehat{\alpha}, \widehat{\beta})$ such that the tuples $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$ of constraint violations corresponding to the candidates $\widehat{\alpha}$ and $\widehat{\beta}$ yield a minimal sum $\widehat{\mathbf{a}} + \widehat{\mathbf{b}}$, as defined in (12). This is of course a special case of the definition (9) of optimality, applied to the set $A \times B$ with summed tuples of constraint violations rather than to the set A with the original tuples of constraint violations.

- (12) $\text{opt}_{<}(A \times B)$ is the set of those candidate pairs $(\widehat{\alpha}, \widehat{\beta})$ in $A \times B$ such that there exists no competing candidate pair (α, β) in $A \times B$ such that the sum $\mathbf{a} + \mathbf{b}$ of the two tuples \mathbf{a} and \mathbf{b} of constraint violations assigned to the two candidates α and β is smaller than the sum $\widehat{\mathbf{a}} + \widehat{\mathbf{b}}$ of the two tuples $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$ of constraint violations assigned to the two candidates $\widehat{\alpha}$ and $\widehat{\beta}$, namely $\mathbf{a} + \mathbf{b} < \widehat{\mathbf{a}} + \widehat{\mathbf{b}}$.

Phonological theories which make use of constraint summation, such as DT and the OPM, assume that the two underlying forms considered here are mapped to the pair of surface candidates $(\widehat{\alpha}, \widehat{\beta})$ which violates the constraints the least, namely which belongs to the optimal set $\text{opt}_{<}(A \times B)$ of candidate pairs with the smallest summed tuple of constraint violations.

Typological innocuousness as a commutativity identity

3.3

Let us take stock. According to the classical implementation of constraint-based phonology reviewed in Subsection 3.1, the order $<$ is used twice. It is used once to assign to the first underlying form a candidate $\widehat{\alpha}$ in the set $\text{opt}_{<}A$ of optimal candidates of A . It is then used again and independently to assign to the second underlying form a candidate $\widehat{\beta}$ in the set $\text{opt}_{<}B$ of optimal candidates of B . According to the constraint summation approach of DT and the OPM reviewed in Subsection 3.2, the order $<$ is instead used only once to assign to the two underlying forms considered simultaneously a candidate pair $(\widehat{\alpha}, \widehat{\beta})$ in the set $\text{opt}_{<}(A \times B)$ of optimal pairs of $A \times B$, where candidate pairs are compared based on the sums of constraint violations of

the two individual candidates, by virtue of the constraint summation assumption (1).

Suppose now that a candidate pair is optimal iff it consists of two optimal candidates, as stated in (13). This means that the two underlying forms considered end up with the same optimal candidates no matter whether we adopt the classical approach or the approach based on constraint summation of DT and the OPM. In other words, the constraint summation assumption (1) made by DT and the OPM would be *typologically innocuous*. And classical constraint-based phonology would thus follow as a special case of DT and the OPM when the constraint set contains no distinctiveness or OP faithfulness constraints.⁷

(13)
Commutativity identity:

$$\underbrace{\text{opt}_{<}(A \times B)}_{\text{constraint summation approach (DT/OPM)}} = \underbrace{\text{opt } A \times \text{opt } B}_{\text{classical approach}}$$

In conclusion, a crucial issue of the formal analysis of phonological theories such as DT and the OPM is whether the identity (13) holds in the general case, for any two candidate sets A and B . In other words, whether the two operations of optimization and product *commute*: by first combining (through \times) candidates from A and B into pairs and then optimizing (through $\text{opt}_{<}$) over candidate pairs relative to the summed constraint violations (as prescribed by the left-hand side, which corresponds to the summation based approach of DT or the OPM) we get the same result that we get by first optimizing (through $\text{opt}_{<}$) within the two separate candidate sets A and B and then combining (through \times) optimal candidates into pairs (as prescribed by the right-hand side, which corresponds to the classical approach in constraint-based phonology).

⁷ As anticipated at the beginning of this section, the discussion extends straightforwardly from the case of only two underlying forms considered here to the case of an arbitrary finite number of underlying forms. Indeed, suppose we have three underlying forms with candidate sets A , B , and C . The commutativity identity that we need to establish in this case is $\text{opt}_{<}(A \times B \times C) = \text{opt}_{<} A \times \text{opt}_{<} B \times \text{opt}_{<} C$. The latter follows by applying (13) twice: once to the two sets $A \cup B$ and C , to ensure that $\text{opt}_{<}(A \times B \times C) = \text{opt}_{<}(A \times B) \times \text{opt}_{<} C$; then again to the two sets A and B , to ensure that $\text{opt}_{<}(A \times B) = \text{opt}_{<} A \times \text{opt}_{<} B$.

*Constraint distinctiveness is not preserved
by constraint summation*

The sum $\mathbf{a} + \mathbf{b}$ of two tuples of constraint violations carries less information than the two individual tuples \mathbf{a} and \mathbf{b} : the summed tuple is computed from the two individual tuples but the individual tuples cannot be univocally reconstructed from the summed tuple. The assumption (1) of constraint summation can thus wipe away potentially crucial information encoded in the individual tuples of constraint violations, imperiling the validity of the commutativity identity (13). To appreciate the problem, let us look at the behavior of constraint distinctiveness under constraint summation.

Suppose that the (classical) constraints C_1, \dots, C_n considered are *distinctive*. This means that any two candidates in the candidate set A and any two candidates in the candidate set B are distinguished by at least one constraint. Equivalently, no two candidates in A and no two candidates in B are assigned identical tuples of constraint violations. Suppose furthermore that the order $<$ over tuples of constraint violations is *total*: any two different tuples are ordered relative to each other. Distinctiveness and totality together ensure that the set $\text{opt}_{<} A$ of optimal candidates of A and the set $\text{opt}_{<} B$ of optimal candidates of B are both singleton sets. Their product $\text{opt}_{<} A \times \text{opt}_{<} B$ on the right-hand side of the commutativity identity (13) therefore consists of a single candidate pair. The commutativity identity thus requires that also the set $\text{opt}_{<} (A \times B)$ of optimal candidate pairs in the product $A \times B$ consists of a single pair. But the assumption that $<$ is a total order does not suffice to ensure that, because $A \times B$ could contain two different pairs of candidates which share the same summed tuple of constraint violations, despite the individual candidates in A and B all having distinct tuples of constraint violations. In other words, distinctiveness can be lost when constraint violations are added together.

As a concrete example, consider the two candidate sets A and B described by the two tableaux (2). The tuples of constraint violations listed there are all distinct. If the order $<$ is total, the two sets $\text{opt}_{<} A$ and $\text{opt}_{<} B$ of optimal candidates are thus each a singleton. And their product $\text{opt}_{<} A \times \text{opt}_{<} B$ on the right-hand side of the commutativity identity (13) thus consists of a single candidate pair. Yet, the product $A \times B$ contains the two different candidate pairs ($[ada]$, $[a^nda]$) and

([aⁿda], [ada]) whose tuples of summed constraint violations are identical, as shown in (11f) and (11h). Suppose that the order $<$ is defined in such a way that this shared summed tuple happens to be minimal relative to the total order $<$. This means that the set $\text{opt}_{<}(A \times B)$ of optimal candidate pairs in $A \times B$ contains both pairs ([ada], [aⁿda]) and ([aⁿda], [ada]). The commutativity identity (13) thus fails, because its right-hand side is a singleton while its left-hand side is not.

These considerations show that we have every reason to expect the commutativity identity (13) to fail in the general case, whereby classical constraint-based phonology cannot be construed as a special case of DT and the OPM, even when the constraint set contains no distinctiveness or OP faithfulness constraints. Can we nonetheless isolate and characterize some special class of orders $<$ among tuples of constraint violations whose special properties validate the commutativity identity (13)? This is the question that we will tackle and solve in the rest of the paper.

4

CONSTRAINT SUMMATION
IS TYPOLOGICALLY INNOCUOUS IN OT:
PRINCE (2015)

In Section 2, we have reviewed some approaches to phonology (such as DT and the OPM) that assume that classical faithfulness and markedness constraints are summed across multiple candidates, as stated in (1). In Section 3, we have formalized the question of the typological innocuousness of constraint summation through the commutativity identity (13). In this section, we review a result by Prince (2015) showing that this commutativity identity indeed holds in OT. In other words, despite constraint summation, theories such as DT and the OPM make the same typological predictions as classical OT when the constraint set only consists of classical constraints and no distinctiveness or OP faithfulness constraints, whereby constraint summation is typologically innocuous. The next section will then extend this result beyond OT.

Prince (2015) focuses on the special case where the order $<$ over tuples of constraint violations is OT's lexicographic order. Let us recall here the explicit definition of this order, that we have already used implicitly in Section 2. We start by linearly ordering or *ranking* the n constraints C_1, C_2, \dots, C_n in some arbitrary way. Without loss of generality, we assume that constraint C_1 is ranked at the top, constraint C_2 is ranked right underneath it, and so on. The inequality $\mathbf{a} < \hat{\mathbf{a}}$ then holds between any two tuples of constraint violations $\mathbf{a} = (a_1, \dots, a_n)$ and $\hat{\mathbf{a}} = (\hat{a}_1, \dots, \hat{a}_n)$ provided there exists some integer k between 1 and n which validates the conditions in (14).

$$(14) \quad \begin{array}{rcl} a_1 & = & \hat{a}_1 \\ & \vdots & \\ a_{k-1} & = & \hat{a}_{k-1} \\ a_k & < & \hat{a}_k \end{array}$$

These conditions say that the $k - 1$ top ranked constraints assign the same number of violations to the two candidates corresponding to the tuples \mathbf{a} and $\hat{\mathbf{a}}$.⁸ And that the k th constraint is then *decisive* because it assigns less violations to the candidate corresponding to the tuple \mathbf{a} than to the candidate corresponding to the tuple $\hat{\mathbf{a}}$. Constraints ranked underneath play no role. In Section 5, we will consider alternative ways of ordering tuples of constraint violations.

Prince's result, rephrased below as Proposition 1, says that no ranking information is lost by summing together constraint violations in the case of OT, in the sense that the commutativity identity (13) holds for any candidate sets. In other words, when the constraint set consists of classical constraints only, theories which use constraint summation (such as DT and the OPM) coincide with classical OT and constraint summation is therefore typologically innocuous.

PROPOSITION 1 (Prince 2015) *The commutativity identity (13) holds for any two candidate sets A and B relative to OT's lexicographic order $<$ corresponding to any constraint ranking: a candidate $\hat{\alpha}$ belongs to the set $\text{opt}_{<} A$ of OT optimal candidates of A and a candidate $\hat{\beta}$ belongs to*

⁸These conditions are interpreted as vacuously true if $k = 1$.

the set $\text{opt}_{<} B$ of OT optimal candidates of B if and only if the candidate pair $(\widehat{\alpha}, \widehat{\beta})$ belongs to the set $\text{opt}_{<}(A \times B)$ of optimal candidate pairs in $A \times B$, when candidate pairs are compared based on summed constraint violations.

4.2

A simple proof of Prince’s result

Prince proves Proposition 1 using a piece of notation specifically tailored to OT, namely *elementary ranking conditions* (ERCs; Prince 2002). But this line of reasoning turns out to be involved, intuitively because the operation of constraint summation does not admit a simple counterpart in the theory of ERCs. Yet, Proposition 1 admits an elementary explanation when we reason directly in terms of violation profiles rather than ERCs. In order to streamline the proof of the proposition, we split the commutativity identity (13) into the two inclusions (15) and consider them separately.

$$(15) \quad \begin{array}{l} \text{a. } \text{opt}_{<}(A \times B) \subseteq \text{opt}_{<} A \times \text{opt}_{<} B \\ \text{b. } \text{opt}_{<}(A \times B) \supseteq \text{opt}_{<} A \times \text{opt}_{<} B \end{array}$$

To establish the inclusion (15a), let us assume by contradiction that it fails. This means that the candidate pair $(\widehat{\alpha}, \widehat{\beta})$ is OT optimal in $A \times B$ but that, say, the candidate $\widehat{\alpha}$ is not OT optimal in A . This contradictory assumption means that there exists a different candidate α in A that beats (has smaller constraint violations than) candidate $\widehat{\alpha}$. In other words, the tuples $\mathbf{a} = (a_1, \dots, a_n)$ and $\widehat{\mathbf{a}} = (\widehat{a}_1, \dots, \widehat{a}_n)$ of constraint violations of the two candidates α and $\widehat{\alpha}$ satisfy the inequality $\mathbf{a} < \widehat{\mathbf{a}}$. This inequality says that there exists $k \in \{1, \dots, n\}$ such that conditions (14) hold. By adding the corresponding components $\widehat{b}_1, \dots, \widehat{b}_{k-1}, \widehat{b}_k$ of the tuple $\widehat{\mathbf{b}}$ of constraint violations of candidate $\widehat{\beta}$ to both sides of the inequalities (14), we obtain (16).

$$(16) \quad \begin{array}{rcl} a_1 + \widehat{b}_1 & = & \widehat{a}_1 + \widehat{b}_1 \\ & \vdots & \\ a_{k-1} + \widehat{b}_{k-1} & = & \widehat{a}_{k-1} + \widehat{b}_{k-1} \\ a_k + \widehat{b}_k & < & \widehat{a}_k + \widehat{b}_k \end{array}$$

Conditions (16) say that $\mathbf{a} + \widehat{\mathbf{b}} < \widehat{\mathbf{a}} + \widehat{\mathbf{b}}$. In other words, the candidate pair $(\alpha, \widehat{\beta})$ beats the candidate pair $(\widehat{\alpha}, \widehat{\beta})$. This conclusion contradicts the assumption that the candidate pair $(\widehat{\alpha}, \widehat{\beta})$ is OT optimal in $A \times B$.

The proof of the reverse inclusion (15b) is analogous. Indeed, let us assume by contradiction that the candidate $\widehat{\alpha}$ is OT optimal in A and that the candidate $\widehat{\beta}$ is OT optimal in B but that the pair $(\widehat{\alpha}, \widehat{\beta})$ is not OT optimal in $A \times B$. This means that there exists a different pair (α, β) in $A \times B$ such that $\mathbf{a} + \mathbf{b} < \widehat{\mathbf{a}} + \widehat{\mathbf{b}}$, where $\mathbf{a}, \mathbf{b}, \widehat{\mathbf{a}}, \widehat{\mathbf{b}}$ are the tuples of constraint violations of the four candidates $\alpha, \beta, \widehat{\alpha}, \widehat{\beta}$. Suppose that $\mathbf{a} \neq \widehat{\mathbf{a}}$. Since the lexicographic order $<$ is total and $\widehat{\alpha}$ is optimal in A , then $\widehat{\mathbf{a}} < \mathbf{a}$. This means that there exists h such that $\widehat{a}_1 = a_1, \dots, \widehat{a}_{h-1} = a_{h-1}, \widehat{a}_h < a_h$. Analogously, suppose that $\mathbf{b} \neq \widehat{\mathbf{b}}$. Again, since $<$ is a total order and $\widehat{\beta}$ is optimal in B , then $\widehat{\mathbf{b}} < \mathbf{b}$. This means that there exists k such that $\widehat{b}_1 = b_1, \dots, \widehat{b}_{k-1} = b_{k-1}, \widehat{b}_k < b_k$. Suppose without loss of generality that $h \geq k$. Thus $\widehat{a}_1 + \widehat{b}_1 = a_1 + b_1, \dots, \widehat{a}_{k-1} + \widehat{b}_{k-1} = a_{k-1} + b_{k-1}, \widehat{a}_k + \widehat{b}_k < a_k + b_k$. This means that $\widehat{\mathbf{a}} + \widehat{\mathbf{b}} < \mathbf{a} + \mathbf{b}$, contradicting the assumption $\mathbf{a} + \mathbf{b} < \widehat{\mathbf{a}} + \widehat{\mathbf{b}}$. The cases where either $\mathbf{a} = \widehat{\mathbf{a}}$ or $\mathbf{b} = \widehat{\mathbf{b}}$ are treated analogously.

Back to the issue of constraint distinctiveness

4.3

Having understood the reasoning behind Prince's Proposition 1, let us now go back to the issue of constraint distinctiveness discussed in Subsection 3.4. We suppose that the candidate sets A and B are distinctive: no two candidates in A and no two candidates in B share the same tuple of constraint violations. Since OT's lexicographic order $<$ is total, the corresponding sets $\text{opt}_{<} A$ and $\text{opt}_{<} B$ of OT optimal candidates are both singletons. Their product $\text{opt}_{<} A \times \text{opt}_{<} B$ thus consists of a single pair. Yet, there can exist two different candidate pairs (α, β) and $(\widehat{\alpha}, \widehat{\beta})$ in $A \times B$ which share the same summed tuple $\mathbf{a} + \mathbf{b} = \widehat{\mathbf{a}} + \widehat{\mathbf{b}}$ of constraint violations, because distinctiveness of the individual candidate sets A and B does not entail distinctiveness of their product $A \times B$ when the constraint violations of a pair of candidates are obtained by summing together the constraint violations of the two individual candidates. The assumption that OT's lexicographic order $<$ is total thus does not suffice to ensure that the set $\text{opt}_{<} (A \times B)$ of OT optimal candidate pairs in $A \times B$ is also a singleton. The commutativity identity (13)

could thus in principle fail, because its right-hand side $\text{opt}_{<} A \times \text{opt}_{<} B$ is a singleton while its left-hand side $\text{opt}_{<} (A \times B)$ is not.

But Prince's Proposition 1 ensures that can actually never happen: the two different candidate pairs (α, β) and $(\hat{\alpha}, \hat{\beta})$ which share the same summed tuple $\mathbf{a} + \mathbf{b} = \hat{\mathbf{a}} + \hat{\mathbf{b}}$ of constraint violations can never belong to the set $\text{opt}_{<} (A \times B)$ of OT optimal candidate pairs in $A \times B$. In fact, let us assume by contradiction that they do. Without loss of generality, we assume that the two candidates α and $\hat{\alpha}$ are different (analogous considerations hold if it is the two candidates β and $\hat{\beta}$ that are different instead). Since the candidate set A is distinctive, the tuples \mathbf{a} and $\hat{\mathbf{a}}$ of constraint violations of the two candidates α and $\hat{\alpha}$ must be different. Since $<$ is total, one of these two tuples is larger than the other relative to $<$. Without loss of generality, we suppose that $\mathbf{a} < \hat{\mathbf{a}}$. Crucially, this assumption $\mathbf{a} < \hat{\mathbf{a}}$ entails that $\mathbf{a} + \hat{\mathbf{b}} < \hat{\mathbf{a}} + \hat{\mathbf{b}}$, by reasoning as above from (14) to (16). In other words, the candidate pair $(\alpha, \hat{\beta})$ beats the candidate pair $(\hat{\alpha}, \hat{\beta})$. This conclusion contradicts the assumption that the candidate pair $(\hat{\alpha}, \hat{\beta})$ is OT optimal in $A \times B$.

5 CONSTRAINT SUMMATION IS TYPOLOGICALLY INNOCUOUS: BEYOND OPTIMALITY THEORY

In Section 3, we have formalized typological innocuousness of the constraint summation assumption (1) used by DT and the OPM through the commutativity identity (13). In Section 4, we have recalled from Prince (2015) that this identity holds in the case of the OT model of constraint interaction. In other words, despite constraint summation, theories such as DT and the OPM make the same typological predictions as classical OT when the constraint set only consists of classical constraints and no distinctiveness or OP faithfulness constraints, whereby constraint summation is typologically innocuous.

The focus on OT so far was motivated by the fact that it is the most widely adopted version of constraint-based phonology, and indeed the one adopted in Flemming's implementation of DT and in McCarthy's implementation of the OPM. Yet, the more recent constraint-based phonological literature (Pater 2009; Potts *et al.* 2010) has advocated variants of OT where optimum selection is based on linear utility

functions, as foreshadowed in Goldsmith (1990, §6.5) and Goldsmith (1991, page 259) and advocated in *Linear OT* (LOT; Keller 2000, 2006) and *Harmonic Grammar* (HG; Legendre *et al.* 1990b,a; Smolensky and Legendre 2006). Does the typological innocuousness of the constraint summation assumption (1) extend beyond OT to these alternative implementations of constraint-based phonology? In other words, is the commutativity identity (13) specific to OT's lexicographic order or does it extend to other ways of ordering tuples of constraint violations? This section addresses this question.

Here is a preview of the core result. In Subsection 4.2, we have used two properties of the lexicographic order to establish the commutativity identity (13) for OT. The first property is that the lexicographic ordering of two tuples of constraint violations is not affected by adding the same quantities to the constraint violations in the two tuples, whereby the inequalities (14) entail those in (16). Subsection 5.1 generalizes this property into the notion of *additive* orders. The second property of the lexicographic order that we have used in Subsection 4.2 to establish the commutativity identity for OT is that it is total. This means that any two tuples of constraint violations which are not ordered (neither is larger than the other) must be identical. Subsection 5.2 generalizes total orders to *weak* orders: tuples which are not ordered need not be identical but must be *equivalent*, namely need to satisfy some generalization of the notion of identity. Subsection 5.4 finally shows that additive weak orders are the minimal structure required by a constraint-based phonological formalism to satisfy the commutativity identity (13) and thus to ensure the typological innocuousness of the constraint summation assumption made by DT and the OPM. The proof of this result relies on some properties of additive weak orders established in Subsection 5.3.

Additive orders

5.1

Throughout this section, we consider an arbitrary *strict order* $<$. This means that $<$ satisfies the following three conditions for any tuples $\mathbf{a}, \mathbf{b}, \mathbf{c}$ of constraint violations: it is *irreflexive*, namely $\mathbf{a} < \mathbf{a}$ never holds; it is *asymmetric*, namely $\mathbf{a} < \mathbf{b}$ and $\mathbf{b} < \mathbf{a}$ never both hold; it is *transitive*, namely $\mathbf{a} < \mathbf{b}$ and $\mathbf{b} < \mathbf{c}$ entail $\mathbf{a} < \mathbf{c}$. Recall from (10)

that $\mathbf{a} + \mathbf{b} = (a_1 + b_1, \dots, a_n + b_n)$ denotes the component-wise sum of two tuples $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n)$ of constraint violations. The implication (17) captures the intuitive idea that, if \mathbf{a} is smaller than \mathbf{b} and if the same quantity \mathbf{c} is added to both, the resulting sum $\mathbf{a} + \mathbf{c}$ ought to be smaller than the sum $\mathbf{b} + \mathbf{c}$. Although intuitive, it is possible to construct orders which fail at this condition (one such example is provided in Subsection 6.3). A strict order $<$ which satisfies condition (17) for any three tuples $\mathbf{a}, \mathbf{b}, \mathbf{c}$ of constraint violations is called *additive* (Anderson and Feil 1988).

$$(17) \quad \begin{array}{l} \text{If: } \mathbf{a} < \mathbf{b}, \\ \text{then: } \mathbf{a} + \mathbf{c} < \mathbf{b} + \mathbf{c}. \end{array}$$

To illustrate, OT's lexicographic order is additive (see Subsection 6.1 below for more details).

Throughout this section, we assume that constraint violations can be either positive or negative integers. In other words, we assume that the order $<$ is defined over arbitrary tuples of integers, not necessarily nonnegative integers.⁹ This assumption effectively means that in the consequent of the additivity condition (17), we can either add to or subtract from the constraint violations listed in the tuples \mathbf{a} and \mathbf{b} . This flexibility will be crucial for some of the reasoning developed in this section, such as the proof of condition (21) below. This assumption does not restrict the scope of our results, because the orders of interest considered in Subsection 6 (such as OT's lexicographic order and HG's order based on linear utility functions) can indeed all be construed as ranging over tuples of positive and negative numbers.

The additivity condition (17) entails the variant in (18) for any four tuples $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ of constraint violations.

$$(18) \quad \begin{array}{l} \text{If: } \mathbf{a} < \mathbf{b} \text{ and } \mathbf{c} < \mathbf{d}, \\ \text{then: } \mathbf{a} + \mathbf{c} < \mathbf{b} + \mathbf{d}. \end{array}$$

In fact, the assumption $\mathbf{a} < \mathbf{b}$ in the antecedent of (18) ensures that $\mathbf{a} + \mathbf{c} < \mathbf{b} + \mathbf{c}$ through the additivity condition (17). Analogously, the assumption $\mathbf{c} < \mathbf{d}$ ensures that $\mathbf{b} + \mathbf{c} < \mathbf{b} + \mathbf{d}$. Finally, the two conditions $\mathbf{a} + \mathbf{c} < \mathbf{b} + \mathbf{c}$ and $\mathbf{b} + \mathbf{c} < \mathbf{b} + \mathbf{d}$ thus obtained ensure the consequent $\mathbf{a} + \mathbf{c} < \mathbf{b} + \mathbf{d}$ of (18), because the order $<$ is transitive.

⁹The additivity assumption (17) thus means that $(\mathbb{Z}^n, +, <)$ is an *ordered group* (Anderson and Feil 1988).

As motivated in Subsection 3.1 (see in particular footnote 6), we allow for the possibility that the strict order $<$ is *partial*, not necessarily total. This means that there can exist two tuples \mathbf{a}, \mathbf{b} of constraint violations such that neither $\mathbf{a} < \mathbf{b}$ nor $\mathbf{b} < \mathbf{a}$. In this case, we say that \mathbf{a} and \mathbf{b} are *incommensurable* (relative to $<$) and we write $\mathbf{a} \sim \mathbf{b}$. In other words, the partial strict order defines a corresponding *incommensurability relation* \sim , as in (19).

(19) $\mathbf{a} \sim \mathbf{b}$ if and only if neither $\mathbf{a} < \mathbf{b}$ nor $\mathbf{b} < \mathbf{a}$.

Since the strict order $<$ is irreflexive, the inequality $\mathbf{a} < \mathbf{a}$ fails for any tuple \mathbf{a} of constraint violations. In other words, any tuple \mathbf{a} is incommensurable with itself and the incommensurability relation \sim is therefore reflexive. Furthermore, the incommensurability relation \sim is obviously symmetric. The strict order $<$ is called *weak* provided the corresponding incommensurability relation \sim is also transitive, namely qualifies as an equivalence relation among tuples of constraint violations (Roberts and Tesman 2005, section 4.2.4). We will see some examples below in Subsection 6.

The intuition behind this definition is that a weak order $<$ orders two incommensurable tuples in the same way relative to any other tuples, in the sense that the implication (20) holds for any three tuples $\mathbf{a}, \mathbf{b}, \mathbf{c}$ of constraint violations.

(20) If: $\mathbf{a} < \mathbf{b}$ and $\mathbf{b} \sim \mathbf{c}$,
 then: $\mathbf{a} < \mathbf{c}$.

In fact, let us assume by contradiction that the consequent $\mathbf{a} < \mathbf{c}$ of (20) fails. This means that either $\mathbf{a} \sim \mathbf{c}$ or $\mathbf{c} < \mathbf{a}$. But $\mathbf{a} \sim \mathbf{c}$ is impossible, because together with the assumption $\mathbf{b} \sim \mathbf{c}$ and the transitivity of \sim , it would entail $\mathbf{a} \sim \mathbf{b}$, contradicting the other assumption $\mathbf{a} < \mathbf{b}$. Analogously, $\mathbf{c} < \mathbf{a}$ is impossible as well, because together with the assumption $\mathbf{a} < \mathbf{b}$ and the transitivity of $<$, it would entail $\mathbf{c} < \mathbf{b}$, contradicting the other assumption $\mathbf{b} \sim \mathbf{c}$.

We now have two assumptions on the strict partial order $<$ over tuples of (positive and negative) constraint violations: that it is additive

and that it is a weak order. Subsection 5.4 will show that these two assumptions are necessary and sufficient to guarantee the commutativity identity (13) and thus to ensure that the constraint summation assumption (1) in DT and the OPM is typologically innocuous. Towards establishing this result, we now take a closer look at the combination of these two assumptions that an order is both additive and weak.

Suppose that the strict order $<$ is additive, in the sense that it satisfies condition (17). Its corresponding incommensurability relation \sim is then additive as well, in the sense that it satisfies the completely analogous condition (21) for any tuples $\mathbf{a}, \mathbf{b}, \mathbf{c}$ of constraint violations.

$$(21) \quad \begin{array}{ll} \text{If:} & \mathbf{a} \sim \mathbf{b}, \\ \text{then:} & \mathbf{a} + \mathbf{c} \sim \mathbf{b} + \mathbf{c}. \end{array}$$

In fact, let us assume by contradiction that the consequent $\mathbf{a} + \mathbf{c} \sim \mathbf{b} + \mathbf{c}$ of (21) fails. This means that either $\mathbf{b} + \mathbf{c} < \mathbf{a} + \mathbf{c}$ or $\mathbf{a} + \mathbf{c} < \mathbf{b} + \mathbf{c}$. For concreteness, let us suppose that the former case $\mathbf{b} + \mathbf{c} < \mathbf{a} + \mathbf{c}$ holds. Adding $-\mathbf{c}$ to both sides (which we are allowed to do, because we are not restricting ourselves to nonnegative constraint violations) yields $\mathbf{b} < \mathbf{a}$, because the order $<$ satisfies the additivity condition (17). This conclusion $\mathbf{b} < \mathbf{a}$ contradicts the assumption $\mathbf{a} \sim \mathbf{b}$.

The reasoning used in Subsection 5.1 to show that the original additivity condition (17) for the order $<$ entails the variant (18) can be rebooted here to show that the additivity condition (21) for the incommensurability relation \sim entails the analogous variant (22) for any four tuples $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ of constraint violations.

$$(22) \quad \begin{array}{ll} \text{If:} & \mathbf{a} \sim \mathbf{b} \text{ and } \mathbf{c} \sim \mathbf{d}, \\ \text{then:} & \mathbf{a} + \mathbf{c} \sim \mathbf{b} + \mathbf{d}. \end{array}$$

In fact, the assumption $\mathbf{a} \sim \mathbf{b}$ in the antecedent of (22) ensures that $\mathbf{a} + \mathbf{c} \sim \mathbf{b} + \mathbf{c}$ through the additivity condition (21). Analogously, the assumption $\mathbf{c} \sim \mathbf{d}$ ensures that $\mathbf{b} + \mathbf{c} \sim \mathbf{b} + \mathbf{d}$. Finally, the two conditions $\mathbf{a} + \mathbf{c} \sim \mathbf{b} + \mathbf{c}$ and $\mathbf{b} + \mathbf{c} \sim \mathbf{b} + \mathbf{d}$ thus obtained ensure the consequent $\mathbf{a} + \mathbf{c} \sim \mathbf{b} + \mathbf{d}$ of (22), because the incommensurability relation \sim is transitive.

Conditions (17)/(18) and (21)/(22) characterize additivity of the order $<$ and of the incommensurability relation \sim *separately*. They entail the following mixed additivity condition (23), which features

the two relations jointly. This condition says that the validity of an inequality is not affected by adding incommensurable elements at both sides.

$$(23) \quad \begin{array}{l} \text{If: } a < b \text{ and } c \sim d, \\ \text{then: } a + c < b + d. \end{array}$$

In fact, the assumption $a < b$ in the antecedent of (23) ensures that $a + c < b + c$ through the additivity condition (17) for the order $<$. Analogously, the assumption $c \sim d$ ensures that $b + c \sim b + d$, through the additivity condition (21) for the incommensurability relation \sim . Finally, the two conditions $a + c < b + c$ and $b + c \sim b + d$ thus obtained ensure the consequent $a + c < b + d$ of (23), because of the condition (20) that incommensurable tuples are ordered alike.

As noted above, since the strict order $<$ is irreflexive, its incommensurability relation \sim is reflexive. This means in particular that $c \sim d$ whenever $c = d$. The mixed additivity condition (23) thus generalizes the original additivity condition (17) for the weak order $<$ from the special case $c = d$ to the more general case $c \sim d$. This generalization makes intuitive sense because the assumption that $<$ is a weak order means that its incommensurability relation \sim is an equivalence relation, namely that \sim generalizes the identity $=$.

We conclude this subsection with the characterization of additive weak orders provided by the following lemma, in terms of the two additivity conditions (17) and (21) or equivalently in terms of the mixed additivity condition (23). The next subsection will use this characterization to establish a connection between additive weak orders and the commutativity identity (13) which was shown to be crucial for the typological innocuousness of the constraint summation assumption made by theories such as DT and the OPM.

LEMMA 1 *A strict (possibly partial) order $<$ is an additive weak order if and only if it satisfies the additivity condition (17) and furthermore its incommensurability relation satisfies the additivity condition (21). Equivalently, if and only if it satisfies the mixed additivity condition (23).*

PROOF We only need to show that the mixed additivity condition (23) entails transitivity of the incommensurability relation \sim . Let us assume by contradiction that \sim is not transitive, namely that $a \sim b$ and $b \sim c$ but $a \not\sim c$. The latter condition $a \not\sim c$ means that either

$a < c$ or $c < a$; for concreteness, we assume that the former case holds. By (23), $a < c$ and $c \sim b$ entail $a + c < c + b$. By (23) again, $a + c < c + b$ and $-c \sim -c$ (which holds because \sim is reflexive, as it is the incommensurability relation of a strict order), entail $a < b$. This conclusion $a < b$ contradicts the hypothesis $a \sim b$. \square

5.4

*The commutativity identity (13)
holds for (and only for) additive weak orders*

This section proves the following Proposition 2, which is the main result of this paper. The “if” statement of the proposition says that additive weak orders provide *sufficient* structure to ensure the commutativity identity (13). Furthermore, the “only if” statement says that additive weak orders provide the *necessary* structure for the commutativity identity (13) to hold. In other words, the constraint summation assumption (1) made by DT and the OPM is typologically innocuous if and only if tuples of constraint violations are compared and optimized relative to an additive weak order.

PROPOSITION 2 *Consider a strict (possibly partial) order $<$ over tuples of constraint violations. The commutativity identity (13) repeated below holds for any two candidate sets A and B if and only if $<$ is an additive weak order.*

$$(13) \quad \boxed{\text{opt}_{<}(A \times B) = \text{opt}_{<} A \times \text{opt}_{<} B}$$

PROOF In order to streamline the proof of the proposition, it useful to split the commutativity identity (13) into the two inclusions (24).

$$(24) \quad \begin{aligned} \text{a. } & \text{opt}_{<}(A \times B) \subseteq \text{opt}_{<} A \times \text{opt}_{<} B \\ \text{b. } & \text{opt}_{<}(A \times B) \supseteq \text{opt}_{<} A \times \text{opt}_{<} B \end{aligned}$$

The proof of the proposition relies on the characterization of additive weak orders provided by Lemma 1 through the additivity condition (17) for the order $<$ and the additivity condition (21) for the incommensurability relation \sim , as summarized in (25). The additivity condition (17) for $<$ suffices to derive the inclusion (24a) while both additivity conditions (17) and (21) for $<$ and \sim are needed to derive

the reverse inclusion (24b). Vice versa, the inclusions (24a) and (24b) each suffice to derive the additivity conditions (17) and (21) for $<$ and \sim , respectively.

- (25) a. $<$ additivity condition (17) \implies inclusion (24a)
 b. $\left. \begin{array}{l} < \text{ additivity condition (17) } \\ \sim \text{ additivity condition (21) } \end{array} \right\} \implies$ inclusion (24b)
 c. $<$ additivity condition (17) \iff 4 inclusion (24a)
 d. \sim additivity condition (21) \iff inclusion (24b)

We start by showing that the additivity condition (17) for $<$ entails the inclusion (24a), as stated in (25a). We consider a candidate pair $(\widehat{\alpha}, \widehat{\beta})$ that belongs to the set $\text{opt}_{<}(A \times B)$ of optimal candidate pairs of $A \times B$. We suppose by contradiction that either the candidate $\widehat{\alpha}$ does not belong to the set $\text{opt}_{<}A$ of optimal candidates of A or the candidate $\widehat{\beta}$ does not belong to the set $\text{opt}_{<}B$ of optimal candidates of B (or both). For concreteness, we assume that the former case holds, namely that $\widehat{\alpha}$ does not belong to the optimal set $\text{opt}_{<}A$. This means in turn that there exists another candidate α of A such that $4a < \widehat{\mathbf{a}}$, where \mathbf{a} and $\widehat{\mathbf{a}}$ are the tuples of constraint violations of the two candidates α and $\widehat{\alpha}$, respectively. Since $<$ satisfies the additivity condition (17), $\mathbf{a} < \widehat{\mathbf{a}}$ entails $\mathbf{a} + \widehat{\mathbf{b}} < \widehat{\mathbf{a}} + \widehat{\mathbf{b}}$, where $\widehat{\mathbf{b}}$ is the tuple of constraint violations of the candidate $\widehat{\beta}$. This inequality $\mathbf{a} + \widehat{\mathbf{b}} < \widehat{\mathbf{a}} + \widehat{\mathbf{b}}$ says that the candidate pair $(\alpha, \widehat{\beta})$ beats the candidate pair $(\widehat{\alpha}, \widehat{\beta})$. This conclusion contradicts the hypothesis that the candidate pair $(\widehat{\alpha}, \widehat{\beta})$ belongs to the set $\text{opt}_{<}(A \times B)$ of optimal candidate pairs of $A \times B$. This reasoning is analogous to the reasoning used in Subsection 4.1 to prove the inclusion (15a).

We show next that the two additivity conditions (17) and (21) – and their corollaries (18), (22), and (23) – entail the other inclusion (24b), as stated in (25b). Consider a candidate $\widehat{\alpha}$ that belongs to the set $\text{opt}_{<}A$ of optimal candidates of A . Consider next a candidate $\widehat{\beta}$ that belongs to the set $\text{opt}_{<}B$ of optimal candidates of B . We assume by contradiction that the candidate pair $(\widehat{\alpha}, \widehat{\beta})$ does not belong to the set $\text{opt}_{<}(A \times B)$ of optimal candidate pairs of $A \times B$. This means that there exists another candidate pair (α, β) in $A \times B$ such that the sum $\mathbf{a} + \mathbf{b}$ of the tuples \mathbf{a} and \mathbf{b} of constraint violations of candidates α and β is smaller than the sum $\widehat{\mathbf{a}} + \widehat{\mathbf{b}}$ of the tuples $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$ of constraint

violations of the candidates $\hat{\alpha}$ and $\hat{\beta}$, namely $\mathbf{a} + \mathbf{b} < \hat{\mathbf{a}} + \hat{\mathbf{b}}$. Since the candidate α belongs to A and the candidate $\hat{\alpha}$ is optimal for A , either $\hat{\mathbf{a}} < \mathbf{a}$ or else $\mathbf{a} \sim \hat{\mathbf{a}}$. Analogously, since the candidate β belongs to B and the candidate $\hat{\beta}$ is optimal for B , either $\hat{\mathbf{b}} < \mathbf{b}$ or else $\mathbf{b} \sim \hat{\mathbf{b}}$. If $\hat{\mathbf{a}} < \mathbf{a}$ and $\hat{\mathbf{b}} < \mathbf{b}$, the additivity condition (18) entails $\hat{\mathbf{a}} + \hat{\mathbf{b}} < \mathbf{a} + \mathbf{b}$, which contradicts the assumption $\mathbf{a} + \mathbf{b} < \hat{\mathbf{a}} + \hat{\mathbf{b}}$. If $\hat{\mathbf{a}} < \mathbf{a}$ and $\hat{\mathbf{b}} \sim \mathbf{b}$ (or if $\hat{\mathbf{a}} \sim \mathbf{a}$ and $\hat{\mathbf{b}} < \mathbf{b}$), the mixed additivity condition (23) entails $\hat{\mathbf{a}} + \hat{\mathbf{b}} < \mathbf{a} + \mathbf{b}$, which again contradicts the assumption $\mathbf{a} + \mathbf{b} < \hat{\mathbf{a}} + \hat{\mathbf{b}}$. Finally, if $\hat{\mathbf{a}} \sim \mathbf{a}$ and $\hat{\mathbf{b}} \sim \mathbf{b}$, the additivity condition (22) for the incommensurability relation entails $\hat{\mathbf{a}} + \hat{\mathbf{b}} \sim \mathbf{a} + \mathbf{b}$, which again contradicts the assumption $\mathbf{a} + \mathbf{b} < \hat{\mathbf{a}} + \hat{\mathbf{b}}$. This reasoning is analogous to the reasoning used in Subsection 4.1 to prove the inclusion (15b).

Turning to the opposite direction, we show now that the inclusion (24a) entails that the order $<$ satisfies the additivity condition (17), as stated in (25c). Thus, we assume that the antecedent $\mathbf{a} < \mathbf{b}$ of the additivity condition (17) holds and we consider an arbitrary third vector \mathbf{c} . We consider a set $A = \{\alpha, \beta\}$ consisting of two candidates α and β whose tuples of constraint violations are \mathbf{a} and \mathbf{b} . Furthermore, we consider a set $B = \{\gamma\}$ consisting of a unique candidate γ whose tuple of constraint violations is \mathbf{c} . The hypothesis $\mathbf{a} < \mathbf{b}$ means that the set $\text{opt}_{<} A$ of optimal candidates of A only consists of the candidate α . Furthermore, the set $\text{opt}_{<} B$ of optimal candidates of B only consists of the candidate γ , because B is a singleton. Hence, the product $\text{opt}_{<} A \times \text{opt}_{<} B$ of the two optimal sets only consists of the pair (α, γ) . Finally, $A \times B = \{(\alpha, \gamma), (\beta, \gamma)\}$. The inclusion (24a) thus says that the set $\text{opt}_{<} (A \times B)$ of optimal candidate pairs of $A \times B$ only consists of the pair (α, γ) and does not contain the other pair (β, γ) . This means in turn that neither $\mathbf{a} + \mathbf{c} \sim \mathbf{b} + \mathbf{c}$ nor $\mathbf{b} + \mathbf{c} < \mathbf{a} + \mathbf{c}$ and thus that $\mathbf{a} + \mathbf{c} < \mathbf{b} + \mathbf{c}$. This conclusion shows that the additivity condition (17) holds, namely that $\mathbf{a} < \mathbf{b}$ entails $\mathbf{a} + \mathbf{c} < \mathbf{b} + \mathbf{c}$.

We conclude by showing that the inclusion (24b) entails that the incommensurability relation \sim satisfies the additivity condition (21), as stated in (25d). Thus, we assume that the antecedent $\mathbf{a} \sim \mathbf{b}$ of the additivity condition (21) holds and we consider an arbitrary third vector \mathbf{c} . We consider the candidate sets $A = \{\alpha, \beta\}$ and $B = \{\gamma\}$ as above, where the three candidates α, β, γ have the tuples of constraint violations $\mathbf{a}, \mathbf{b}, \mathbf{c}$, respectively. The incommensurability assumption $\mathbf{a} \sim \mathbf{b}$

says that $\text{opt}_{<} A = A$. Furthermore, $\text{opt}_{<} B = B$, because B is a singleton. Hence, $\text{opt}_{<} A \times \text{opt}_{<} B = A \times B$. The inclusion (24b) thus says that $A \times B = \text{opt}_{<} (A \times B)$. In other words, both candidate pairs (α, β) and (α, γ) of the set $A \times B$ actually belong to the optimal set $\text{opt}_{<} (A \times B)$. This means in turn that the tuples of constraint violations of these two candidate pairs (α, β) and (α, γ) are incommensurable, namely that $\mathbf{a} + \mathbf{c} \sim \mathbf{b} + \mathbf{c}$. This conclusion shows that the additivity condition (21) holds, namely that $\mathbf{a} \sim \mathbf{b}$ entails $\mathbf{a} + \mathbf{c} \sim \mathbf{b} + \mathbf{c}$. \square

APPLICATIONS

6

This section re-derives Prince's proposition 1 that OT satisfies the commutativity identity (13) as a special case of Proposition 2 obtained in the preceding section. Furthermore, it shows that the commutativity identity extends to constraint-based phonological frameworks that order tuples of constraint violations based on additive utility functions. It follows in particular (see Proposition 3) that the commutativity identity holds for HG. In other words, the typological innocuousness of the constraint summation assumption made by DT and the OPM extends from the OT to the HG mode of constraint interaction.

Re-deriving Prince's result for OT

6.1

Any strict order which is *total* (namely defined for any pair of different tuples of constraint violations) is in particular a weak order. In fact, totality means that two tuples of constraint violations are incommensurable only if they are identical, whereby the incommensurability relation \sim coincides with the identity and it is therefore transitive. Proposition 2 thus ensures that the commutativity identity (13) crucial for DT and the OPM holds whenever grammatical optimization is relative to a total additive strict order.

As our first application of Proposition 2, we can now derive anew Prince's Proposition 1 for OT. In fact, let $<$ be OT's lexicographic order corresponding to some ranking of the n constraints. We assume without loss of generality that C_1 is ranked at the top, followed by C_2 ,

and so on. As reviewed in Subsection 4.1, the condition $\mathbf{a} < \mathbf{b}$ then holds for two tuples $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n)$ of constraint violations if and only if conditions (26) hold for some $k \in \{1, \dots, n\}$.

$$(26) \quad \begin{array}{rcl} a_1 & = & b_1 \\ & \vdots & \\ a_k & = & b_k \\ a_{k+1} & < & b_{k+1} \end{array}$$

The lexicographic order $<$ is total, namely defined for any two different tuples of constraint violations. Furthermore, it is additive, namely it satisfies the implication (17): the assumption $\mathbf{a} < \mathbf{b}$ that (26) holds entails the conclusion $\mathbf{a} + \mathbf{c} < \mathbf{b} + \mathbf{c}$ that (27) holds as well.

$$(27) \quad \begin{array}{rcl} a_1 + c_1 & = & b_1 + c_1 \\ & \vdots & \\ a_k + c_k & = & b_k + c_k \\ a_{k+1} + c_{k+1} & < & b_{k+1} + c_{k+1} \end{array}$$

Prince's Proposition 1 for OT thus follows as a special case of Proposition 2: the commutativity identity (13) holds in the case of OT because grammatical optimization in OT is computed relative to the lexicographic order which is additive and total.

6.2

Extension to HG

To explore further applications of Proposition 2, we consider a *utility function* U which assigns to each tuple \mathbf{a} of constraint violations a number $U(\mathbf{a})$. We can then order the tuples of constraint violations based on their utility, with smaller tuples corresponding to a smaller utility, as in (28).

$$(28) \quad \mathbf{a} < \mathbf{b} \text{ if and only if } U(\mathbf{a}) < U(\mathbf{b}).$$

The resulting relation $<$ is obviously a strict order. It is partial, because tuples of constraint violations which achieve the same utility are incommensurable. Furthermore, it is a weak order, because the corresponding incommensurability relation \sim described in (29) is obviously transitive.

(29) $\mathbf{a} \sim \mathbf{b}$ if and only if $U(\mathbf{a}) = U(\mathbf{b})$.

Suppose that the utility function U is *additive*, namely that the identity $U(\mathbf{a} + \mathbf{b}) = U(\mathbf{a}) + U(\mathbf{b})$ holds for any tuples \mathbf{a}, \mathbf{b} of constraint violations. In this case, the corresponding weak strict order $<$ satisfies the additivity condition (17). In fact, the assumption $\mathbf{a} < \mathbf{b}$ of this additivity condition means that $U(\mathbf{a}) < U(\mathbf{b})$. Hence $U(\mathbf{a}) + U(\mathbf{c}) < U(\mathbf{b}) + U(\mathbf{c})$. By additivity, this means $U(\mathbf{a} + \mathbf{c}) < U(\mathbf{b} + \mathbf{c})$, whereby $\mathbf{a} + \mathbf{c} < \mathbf{b} + \mathbf{c}$. Proposition 2 thus ensures that the commutativity identity (13) crucial for DT and the OPM holds whenever grammatical optimization is relative to the order induced by an additive utility function.

Taking advantage of the fact that constraint violations are integers, Magri (2020) shows (through a simple twist of the Fundamental Theorem of Linear Algebra; Strang 2006, Section 2.6) that for any additive utility function U , there exist a *weight vector* $\mathbf{w} = (w_1, \dots, w_n)$ such that the utility $U(\mathbf{a})$ of any tuple of integer constraint violations can be described as the weighted sum of the constraint violations collected in the tuple \mathbf{a} , namely $U(\mathbf{a}) = \sum_{i=1}^n a_i w_i$. In other words, the partial strict order corresponding to an additive utility function in the sense of (28) yields the HG model of grammatical optimization (Legendre *et al.* 1990b,a; Smolensky and Legendre 2006; Pater 2009; Potts *et al.* 2010). Proposition 2 thus ensures that the commutativity identity (13) crucial for DT and the OPM extends from OT to HG, as stated by the following proposition.

PROPOSITION 3 *The commutativity identity (13) holds for any two candidate sets A and B relative to HG's order $<$ corresponding to any constraint weighting: a candidate $\hat{\alpha}$ belongs to the set $\text{opt}_{<} A$ of optimal candidates of A relative to the HG order $<$ corresponding to that weighting and a candidate $\hat{\beta}$ belongs to the set $\text{opt}_{<} B$ of optimal candidates of B if and only if the candidate pair $(\hat{\alpha}, \hat{\beta})$ belongs to the set $\text{opt}_{<} (A \times B)$ of optimal candidate pairs in $A \times B$, when candidate pairs are compared based on summed constraint violations.*

When the commutativity identity fails

6.3

Crucially, Proposition 2 provides not only a sufficient but also a necessary condition for the commutativity identity (13) to hold. Thus,

this proposition can be used not only to verify that the commutativity identity holds, as we have done so far, but also to disprove that it does. To illustrate, suppose that there are only $n = 2$ constraints and consider the *quadratic* utility function U defined as in (30) for any pair $\mathbf{a} = (a_1, a_2)$ of constraint violations.

$$(30) \quad U(\mathbf{a}) = a_1^2 + a_2^2$$

The corresponding relation $<$ as in (28) is a weak partial strict order. Yet, it does not satisfy the additivity condition (17). In fact, consider for instance $\mathbf{a} = (2, 2)$, $\mathbf{b} = (0, 3)$ and $\mathbf{c} = (4, 4)$. In this case, $\mathbf{a} < \mathbf{b}$ (because $U(\mathbf{a}) = 2^2 + 2^2 = 8$ while $U(\mathbf{b}) = 0^2 + 3^2 = 9$). But $\mathbf{b} + \mathbf{c} < \mathbf{a} + \mathbf{c}$ (because $U(\mathbf{a} + \mathbf{c}) = 6^2 + 6^2 = 72$ while $U(\mathbf{b} + \mathbf{c}) = 4^2 + 7^2 = 65$). Proposition 2 therefore ensures that the commutativity identity (13) crucial for DT and the OPM fails when constraint violations are optimized relative to the order induced by the quadratic utility function (30).

Usually in the constraint-based phonological literature (starting with Prince and Smolensky 1993/2004), each underlying form comes with a preassigned set of candidate surface realizations. Each of these candidates is represented as a tuple of constraint violations. These tuples are compared according to some strict (possibly partial) order $<$ that extends the notion of “being smaller than” from single numbers to tuples of numbers. The optimal candidate for a given underlying form is the one which violates the constraints the least, namely the one with the smallest tuple of constraint violations.

DT (Flemming 2002, 2004, 2008) enriches the classical constraint set with distinctiveness constraints. Furthermore, approaches to paradigm uniformity effects such as the OPM (Kenstowicz 1997; McCarthy 2005) enrich the classical constraint set with OP faithfulness constraints. Crucially, classical (faithfulness and markedness) constraints evaluate a single candidate surface form at a time while distinctiveness and OP faithfulness constraints evaluate multiple candidate surface forms simultaneously, relative to their contrastiveness

and their similarity, respectively. As a consequence, the classical constraints need to be “lifted” from a single candidate to multiple candidates. A reasonable way to do that is to sum their violations across multiple candidates.

Does this assumption of constraint summation made by DT and the OPM make sense? We have formulated this question as follows. Suppose that we restrict ourselves to a constraint set which includes no distinctiveness or OP faithfulness constraints but only classical (markedness and faithfulness) constraints. In this case, can we guarantee that the typological predictions of DT and the OPM coincide with those of the classical theory, despite constraint summation? In other words, is the assumption of constraint summation made by DT and the OPM typologically innocuous?

This paper has shown that constraint summation is indeed typologically innocuous if and only if constraint optimization is performed relative to an order $<$ of tuples of constraint violations which is additive and weak. In other words, additive weak orders provide the “minimal structure” (to go back to Talagrand’s admonition in the quote at the beginning of the paper) for typological innocuousness to hold. This technical condition on grammatical optimization is verified for instance in the case of OT and HG. Our result extends and systematizes an earlier independent result for OT obtained by Prince (2015). Our result provides a solid foundation for theories such as DT and the OPM which make use of constraint summation, for a large class of modes of constraint interaction.

REFERENCES

- Adam ALBRIGHT (2010), Base-driven leveling in Yiddish verb paradigms, *Natural Language & Linguistic Theory*, 28(3):475–537.
- Marlow ANDERSON and Todd FEIL (1988), *Lattice ordered groups: an introduction*, D. Reidel Publishing Company, Dordrecht.
- Laure BENUA (1997), *Transderivational identity: phonological relations between words*, Ph.D. thesis, University of Massachusetts, Amherst.
- Juliette BLEVINS (2004), *Evolutionary phonology: the emergence of sound patterns*, Cambridge University Press, Cambridge.

- Paul BOERSMA and Silke HAMANN (2008), The evolution of auditory dispersion in bidirectional constraint grammars, *Phonology*, 25:217–270.
- Jon P. DAYLEY (1989), *Tümpisa (Panamint) Shoshone grammar*, University of California Press, Berkeley.
- Edward FLEMMING (2002), *Auditory representations in phonology*, Routledge.
- Edward FLEMMING (2004), Contrast and perceptual distinctiveness, in Bruce HAYES, Robert KIRCHNER, and Donca STERIADE, editors, *Phonetically-based phonology*, pp. 232–276, Cambridge University Press, Cambridge.
- Edward FLEMMING (2008), The realized input, unpublished manuscript, MIT.
- Edward FLEMMING (2017a), Dispersion effects in phonology, <https://sites.google.com/site/parisseminarcalt/home/edwardscourse>, handouts for a three-day course in Paris.
- Edward FLEMMING (2017b), Dispersion Theory and phonology, in Mark ARONOFF, editor, *The Oxford research encyclopedia of linguistics*, Oxford University Press, Oxford.
- John A. GOLDSMITH (1990), *Autosegmental and metrical phonology*, Basil Blackwell, Oxford.
- John A. GOLDSMITH (1991), Phonology as an intelligent system, in Donna Jo NAPOLI and Judy KEGL, editors, *Bridges between psychology and linguistics: a Swarthmore festschrift for Lila Gleitman*, pp. 247–267, Lawrence Erlbaum, Mahwah, NJ.
- Gregory K. IVERSON and Joseph C. SALMONS (1996), Mixtec prenasalization as hyper-voicing, *International Journal of American Linguistics*, 62:165–175.
- Frank KELLER (2000), *Gradience in Grammar. Experimental and computational aspects of degrees of grammaticality*, Ph.D. thesis, University of Edinburgh.
- Frank KELLER (2006), Linear Optimality Theory as a model of gradience in grammar, in Gisbert FANSELOW, Caroline FÉRY, Ralph VOGEL, and Matthias SCHLESEWSKY, editors, *Gradience in grammar: generative perspectives*, pp. 270–287, Oxford University Press, Oxford.
- Michael KENSTOWICZ (1997), Base identity and uniform exponence: alternatives to cyclicity, in Jacques DURAND and Bernard LAKS, editors, *Current trends in phonology: models and methods*, pp. 363–394, Salford: University of Salford.
- Géraldine LEGENDRE, Yoshiro MIYATA, and Paul SMOLENSKY (1990a), Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: an application, in Morton Ann GERNSBACHER and Sharon J. DERRY, editors, *Proceedings of the 12th annual conference of the Cognitive Science Society*, pp. 884–891, Lawrence Erlbaum Associates, Hillsdale, NJ.

- Géraldine LEGENDRE, Yoshiro MIYATA, and Paul SMOLENSKY (1990b), Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: theoretical foundations, in Morton Ann GERNSBACHER and Sharon J. DERRY, editors, *Proceedings of the 12th annual conference of the Cognitive Science Society*, pp. 388–395, Lawrence Erlbaum, Hillsdale, NJ.
- Giorgio MAGRI (2020), A principled derivation of Harmonic Grammar, in Allyson ETTINGER, Gaja JAROSZ, and Max NELSON, editors, *Proceedings of the third meeting of the Society for Computation in Linguistics*, Association for Computational Linguistics.
- John J. MCCARTHY (2005), Optimal paradigms, in Laura J. DOWNING, T. Alan HALL, and Renate RAFFELSIEFEN, editors, *Paradigms in phonological theory*, Oxford University Press, Oxford.
- John J. MCCARTHY and Alan PRINCE (1995), Faithfulness and reduplicative identity, in Jill BECKMAN, Suzanne URBANCZYK, and Laura WALSH DICKEY, editors, *University of Massachusetts occasional papers in linguistics 18: papers in Optimality Theory*, pp. 249–384, GLSA, Amherst.
- John J. OHALA (1983), The origin of sound patterns in vocal tract constraints, in Peter F. MACNEILAGE, editor, *The production of speech*, pp. 189–216, Springer-Verlag, New York.
- Joe PATER (2009), Weighted constraints in generative linguistics, *Cognitive Science*, 33:999–1035.
- Christopher POTTS, Joe PATER, Karen JESNEY, Rajesh BHATT, and Michael BECKER (2010), Harmonic Grammar with linear programming: from linear systems to linguistic typology, *Phonology*, 27(1):1–41.
- Alan PRINCE (2002), Entailed ranking arguments, <http://roa.rutgers.edu/files/500-0202/500-0202-PRINCE-0-1.PDF>, manuscript (Rutgers University). Available from the Rutgers Optimality Archive as ROA 500.
- Alan PRINCE (2015), One tableau suffices, http://roa.rutgers.edu/content/article/files/1453_alan_prince_4.pdf, manuscript (Rutgers University). Available from the Rutgers Optimality Archive as ROA 1250.
- Alan PRINCE and Paul SMOLENSKY (1993/2004), *Optimality Theory: Constraint Interaction in generative grammar*, Blackwell, Oxford, <http://roa.rutgers.edu>, original version, Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, April 1993. Available from the Rutgers Optimality Archive as ROA 537.
- Fred S. ROBERTS and Barry TESMAN (2005), *Applied combinatorics*, Taylor and Francis Group.

Paul SMOLENSKY and Géraldine LEGENDRE (2006), *The Harmonic Mind*, MIT Press, Cambridge, MA.

Juliet STANTON (2017), *Constraints on the distribution of nasal-stop sequences: an argument for contrast*, Ph.D. thesis, MIT.

Gilbert STRANG (2006), *Linear Algebra and its applications*, Thomson Brooks/Cole.

Michel TALAGRAND (2014), *Upper and lower bounds for stochastic processes*, Springer-Verlag, Berlin, Heidelberg.

Bernard TRANEL (1987), *The sounds of French: An introduction*, Cambridge University Press.

Giorgio Magri

Ⓘ 0000-0003-0350-9235
magrigrg@gmail.com

SFL (CNRS, University of Paris 8,
UPL)

Benjamin Storme

Ⓘ 0000-0002-2560-2265
benjastorme@hotmail.com

University of Lausanne

Giorgio Magri and Benjamin Storme (2020), *Constraint summation in phonological theory*, *Journal of Language Modelling*, 8(2):251–294

Ⓙ <https://dx.doi.org/10.15398/jlm.v8i2.216>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

ⒸⒻ <http://creativecommons.org/licenses/by/4.0/>

Word prediction in computational historical linguistics

Peter Dekker¹ and Willem Zuidema²

¹ Vrije Universiteit Brussel

² University of Amsterdam

ABSTRACT

In this paper, we investigate how the prediction paradigm from machine learning and Natural Language Processing (NLP) can be put to use in computational historical linguistics. We propose *word prediction* as an intermediate task, where the forms of unseen words in some target language are predicted from the forms of the corresponding words in a source language. Word prediction allows us to develop algorithms for phylogenetic tree reconstruction, sound correspondence identification and cognate detection, in ways close to attested methods for linguistic reconstruction. We will discuss different factors, such as data representation and the choice of machine learning model, that have to be taken into account when applying prediction methods in historical linguistics. We present our own implementations and evaluate them on different tasks in historical linguistics.

Keywords:
computational
historical
linguistics,
machine learning,
deep learning

INTRODUCTION

1

How are the languages of the world related and how have they evolved? This is the central question in one of the oldest linguistic disciplines: *historical linguistics*. In this paper, we aim to contribute to

answering these questions using some of the newest methods: computational modelling, machine learning and big data.¹

Our work can be seen as part of what has been called the *quantitative turn in historical linguistics*: computational methods have been applied to automate parts of the workflow of historical linguistics (Jäger and List 2016), which, in part, has become possible due to the increased availability of digital datasets (the new Cross Linguistic Data Formats initiative proposes new standards for a unified representation of cross-linguistic data, Forkel *et al.* (2018), enabling further expansion and connection of datasets in the coming years).

Such large datasets provide new possibilities, but are at the same time too large to be processed by human experts. Research in computational historical linguistics has therefore attempted to automate several tasks in historical linguistics. Different approaches have been applied to *cognate detection* – the task to detect ancestrally related words (cognates) in different languages – (Inkpen *et al.* 2005; List 2012; Rama 2016; Jäger *et al.* 2017; Dellert 2018), inference of sound correspondences (Hruschka *et al.* 2015), protoform reconstruction (Bouchard-Côté *et al.* 2013) and phylogenetic tree reconstruction (Jäger 2015; Chang *et al.* 2015).

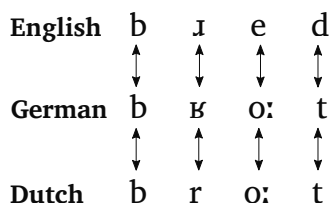
These computational methods have thus opened up many new research directions, and, arguably, provide better replicability than manual methods because of the inherent necessity to specify formal guidelines (Jäger 2019). In recent years, studies in computational historical linguistics have drawn much attention, but also sparked much controversy. Examples are Gray and Atkinson (2003), which charted the age of Indo-European languages, and Bouckaert *et al.* (2012), which proposed to map the Indo-European homeland to Anatolia.

1.1

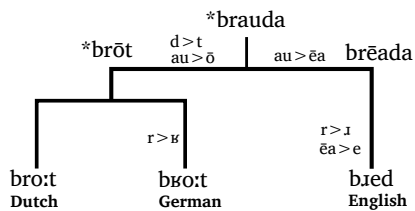
The comparative method

The relation between computational methods and more traditional methods is, however, not always straightforward, and differs in more dimensions than just mathematical formalization. Some computational methods stay conceptually closer to the standard methodology

¹This paper is based on the first author's unpublished MSc thesis (Dekker 2018).



(a) Phenotypic: comparisons are made between the phonemes in aligned word forms



(b) Genotypic: word forms are compared based on a model of the complete ancestral history of the languages, revolving around sound correspondences

Figure 1: Schematic visualization of the difference between phenotypic and genotypic methods, when comparing the English, German and Dutch word forms for the concept *bread*

in historical linguistics: the ‘comparative method’ (Clackson 2007). Linguists applying the comparative method typically make use of a fixed, basic vocabulary, look at the relevant phonetic forms, and focus on *cognates* (words that are ancestrally related). Based on these cognates, they can then identify sound correspondences and, using regular sound correspondences as criterion for descent, reconstruct the ancestral tree of a particular language family, distinguishing between genetic relationships and borrowing.

Although computational methods usually also employ some form of the comparative method, they mostly focus on what can be called *phenotypic* similarity rather than *genotypic* similarity (Lass 1997; List 2012). *Genotypic* methods compare languages based on the language-specific *regular sound correspondences* that can be established between the languages. *Phenotypic* methods compare languages based on the surface forms of words. When comparing words based on surface similarity, it is more difficult to detect the ancestral relatedness of words which underwent much phonetic change and it is more challenging to detect borrowings. Figure 1 shows the difference between phenotypic and genotypic methods schematically.

Genotypic methods are thus preferable to reliably determine ancestral relationship. However, many genotypic methods, like the successful Bayesian MCMC methods for phylogenetic reconstruction, require cognate judgments as input. These cognate judgments have to be performed by human experts, re-introducing human labour, and therefore limiting the amount of data that can be processed. Alterna-

tively, cognate judgments can be automatically inferred via cognate detection methods that are currently developed, but this adds another step to the reconstruction process, with its own inaccuracies.

1.2

Word prediction

An ideal computational method in historical linguistics would automate as much of the process as possible, while still staying close to the comparative method. In this paper, we investigate the usefulness of *word prediction* as an intermediate task that may allow us to arrive at computational methods in historical linguistics. The use of word prediction in historical linguistics was first proposed in the first author's master's thesis (Dekker 2018) and independently by Ciobanu and Dinu (2018), followed by recent approaches (List 2019a; Meloni *et al.* 2019; Cathcart and Wandl 2020; Cathcart and Rama 2020; Fourrier and Sagot 2020a). Word prediction is a methodology that enables the use of surface word forms as data (like phenotypic methods), while still capturing the genetic signal through sound correspondences (like genotypic methods), thus allowing for reliable reconstructions of language relationship based on large amounts of data.

Word prediction allows us to rephrase the reconstruction of language ancestry as a machine learning problem. A machine learning model is trained on pairs of phonetic word forms ($w_{c,A}, w_{c,B}$) denoting the same concept c in two languages A and B . By learning the sound correspondences between the two languages, the model can then predict, for a concept d , the unseen word form $w_{d,B}$, given a word form $w_{d,A}$. Based on the training data, the model learns sound correspondences between the two languages. Therefore, the error between the predicted word and the target word in the test set, for a given source word, provides a distance between source word and target word which is informed by sound correspondences. This contrasts with directly comparing the source and target word – as phenotypic methods do – which yields distances not informed by sound correspondences.

In the rest of this paper, we investigate what is required to successfully apply word prediction in the area of historical linguistics. Our main research questions are the following:

1. What are *suitable data* from historical linguistics and what is a good *representation of input data*, in order to be processed by a machine learning algorithm for the word prediction task?
2. *Which machine learning model* can be used to perform word prediction? A model should be able to learn the relationship between consecutive phonemes (sounds) in a word, and the sound correspondences between source and target word.

In the next sections, we will try to answer these questions, where possible by evaluating multiple alternative solutions. We would like to provide general lessons on factors enabling the use of word prediction in historical linguistics. After these sections, we will develop our own models, based on the answers to the questions we find, and report on the results on different applications in historical linguistics. We end the paper with a discussion of related work and some reflections on the potential and the limitations of word prediction.

DATA 2

Our first question is: which data, and in which representation, are suitable for word prediction? We will first describe which type of data is suitable for this task, and then review different encodings to represent the data.

Datasets 2.1

Data from many linguistic levels can be used to study language change, including lexical, phonetic, morphological and syntactic data. Using word forms (in orthographic or phonetic representation) seems suitable for the prediction task. There are many training examples (words) available per language and the prediction algorithm can generalize over the relations between phonemes. Word forms also have a lower probability of being borrowed or being similar by chance than syntactic data (Greenhill *et al.* 2017). The benefit of phonetic word forms

over orthographic forms is that phonetic forms stay closer to the actual use of language by speakers. Word forms in orthographic representation depend on conventions: the same sound can be described by different letters in different languages.

Ciobanu and Dinu (2018) evaluate their model on three different datasets of orthographic forms for multiple Romance languages (Spanish, Italian, Portuguese, French, Romanian, Latin). In all datasets, the word forms have been grouped into cognate sets: this is needed because the model takes pairs of cognate words as input. The datasets are taken from Bouchard-Côté *et al.* (2007) (585 cognate sets), Reinheimer Ripeanu (2001) (1102 cognate sets) and Ciobanu and Dinu (2014) (3218 cognate sets). Meloni *et al.* (2019) use the Romance cognate dataset from Ciobanu and Dinu (2014) as a basis and augment it with word forms from Wiktionary, arriving at a total of 8799 cognate sets. The authors perform experiments on both orthographic and phonetic word forms. The phonetic word forms are acquired by running a computational transcription library on the orthographic word forms. Cathcart and Wandl (2020) base their dataset on an etymological dictionary by Derksen (2007), and extract Slavic proto-words and their accompanying (cognate) contemporary words in 13 Slavic languages, yielding a dataset of 11 400 forms. Fourrier and Sagot (2020a) use phonetic cognate data from Latin, Spanish and Italian, originating from an etymological database (Fourrier and Sagot 2020b). Fourrier (2020) applies the same workflow, but uses data from Polish, Czech, Lithuanian and Italian.

In our own experiments, we use the NorthEuraLex dataset (Dellert *et al.* 2019),² which consists of phonetic word forms for 1016 concepts in 107 languages in Northern Eurasia. The languages in the dataset belong to many language families (among others Uralic, Indo-European, Turkic and Mongolic), so the number of cognate sets has to be calculated per language family. The size of the dataset is therefore not directly comparable to the size of datasets in Ciobanu and Dinu (2018) and Meloni *et al.* (2019). We use a larger dataset than generally used in historical linguistics. Usually, only basic vocabulary (e.g. kinship terms, body parts) is taken into account because this vocabulary is

² Available for download from <http://northeuralex.org/>. We used the 0.9 release. As of the 0.9.2 release, the dataset contains 30 more languages.

least prone to borrowing (Campbell 2013, p. 352). However, machine learning algorithms need a large number of examples to train on and a meaningful number of examples to evaluate the algorithm. We hope that increasing the performance of the algorithm by using enough training examples compensates for the possible performance decrease caused by borrowing. We use a version of the dataset which is formatted in the *ASJPcode* alphabet (Brown *et al.* 2008). *ASJPcode* consists of 41 sound classes, considerably less than the number of IPA phonemes, reducing the complexity of the prediction problem. For clarity, in this paper, we converted all *ASJP* forms to IPA using the *pyclts* library (Anderson *et al.* 2018).³ As *ASJP* characters represent broader classes of phonetic features than IPA phonemes, the shown IPA phonemes may differ from the original phonemes used in the words.

There can be multiple word forms for a concept in one language. Per language pair, we create word pairs by taking the Cartesian product of all alternative word forms for one concept in both languages. No word pairs are created across concepts. For example, if there are 2 alternative word forms for a concept in language A, and 3 alternative forms for that concept in language B, this yields a total of 6 word pairs. We then split the dataset into a training set (80%), development set (10%) and test set (10%). The training and test set should be separated, so the model predicts on different data than it learned from. The development set is used to tune model parameters (see Section 4).

Data representation

2.2

To enable a machine learning algorithm to process the phonetic data, every phoneme has to be encoded as a numerical vector. We will consider three types of encoding: *one-hot*, *phonetic* and *embedding* encoding.

Ciobanu and Dinu (2018) do not describe their encoding of the data. They do however perform a number of pre-processing steps. First, the word forms of a word pair are aligned using Needleman-Wunsch alignment (Needleman and Wunsch 1970). Subsequently, characters in the output word which remain the same as in the input words, are represented by a special character.

³<https://github.com/cldf-clts/pyclts>

2.2.1

One-hot

In *one-hot encoding*, every phoneme is represented by a vector of length $n_{characters}$, with a 1 at the position which corresponds to the current character, and 0 at all other positions. No qualitative information about the phoneme is stored. Cathcart and Wandl (2020) encode every phoneme of an input word using one-hot encoding, but extends this with a learned embedding per language. Table 1 gives an example of a one-hot feature matrix.

Table 1:
Example of feature matrix for one-hot encoding, for an alphabet consisting of four phonemes. Every phoneme is represented by one feature that is turned on, that feature is unique for that phoneme

IPA				
p	1	0	0	0
b	0	1	0	0
f	0	0	1	0
v	0	0	0	1

2.2.2

Phonetic

In *phonetic encoding*, a phoneme is encoded as a vector of its phonetic features (e.g. back, bilabial, voiced), enabling the model to generalize observed sound changes across different phonemes. Rama (2016), using a neural network approach to cognate detection, shows that a phonetic representation yields better performance than one-hot encoding for some datasets. In our model, we use the phonetic feature matrix for ASJP tokens from Brown *et al.* (2008), formatted as a binary feature matrix by Rama (2016). As mentioned, in this paper, we use IPA to denote the ASJP tokens. Table 2 shows an example of a phonetic feature matrix.

Table 2:
Example of feature matrix for phonetic encoding: every phoneme can have multiple features turned on

IPA	Voiced	Labial	Denta	Alveolar	...
p	0	1	0	0	...
b	1	1	0	0	...
f	0	1	1	0	...
v	1	1	1	0	...
m	1	1	0	0	...
θ	1	0	1	0	...

A third type of encoding that we will consider is the *embedding* encoding, where a linguistic item is encoded using the distribution of items appearing in its context. Most well known are *word embeddings*, which have successfully been applied in many NLP tasks (e.g., Mikolov *et al.* 2013; Pennington *et al.* 2014). The assumption is that “you shall know a word by the company it keeps” (Firth 1957). If two words have a similar embedding vector, they usually appear in the same context and can thus relatively easily be interchanged. But the idea of using embeddings that reflect the context of a linguistic item can also be applied analyzing smaller units than the word level. For instance, it has also successfully been applied in NLP by introducing character-based language models (Kim *et al.* 2016).

In computational historical linguistics, Rama and List (2019) used skip-grams, which capture the context of a phoneme, to perform fast cognate detection. Meloni *et al.* (2019) use an embedding layer in their neural network, which learns an embedding vector of size 100 for every phoneme, from the data. The embedding consists of a language-specific and a language-dependent part. Cathcart and Wandl (2020) use one-hot encoding for phonemes of the input word, and concatenate this with a trained embedding per language.

Similarly, we propose to encode a phoneme as a vector of the phonemes occurring in its context. The same interchangeability of word embeddings is assumed: if two phoneme vectors are similar, they appear in a similar context. This corresponds to language-specific rules in *phonotactics* (the study of the combination of phonemes), which specify that a certain class of phonemes (e.g. approximant) can follow a certain other class (e.g. voiceless fricative). It can be expected that embeddings of phonemes inside a certain class are more likely to be similar to each other than to phonemes in other classes. In some respects, the embedding encoding learns the same feature matrix as the *phonetic* encoding, but inferred from the data, and with more emphasis on language-specific phonotactics.

In our experiments, we also use embedding coding, but do not apply high-dimensional learned embeddings as do Meloni *et al.* (2019). Instead, we want to put more emphasis on the direct neighbours of a phoneme, as most phonotactic rules describe these relations.

Table 3:
 Example of feature matrix
 for embedding encoding:
 every phoneme is represented
 by an array of floating point values,
 which correspond to the probabilities
 that other phonemes occur
 before or after this phoneme.
 The values in a row sum to 1

IPA	START	i LEFT	S LEFT	p RIGHT	...
ə	0.004	0.003	0.001	0.002	...
a	0.024	0.000	0.000	0.003	...
e	0.050	0.002	0.000	0.012	...
b	0.388	0.000	0.000	0.004	...
p	0.152	0.039	0.000	0.000	...

We create language-specific embedding encodings from the whole NorthEuraLex corpus. For every phoneme, the preceding and following phonemes, for all occurrences of the phoneme in the corpus, are counted. Position is taken into account, i.e., an /a/ appearing before a certain phoneme is counted separately from an /a/ appearing after a certain phoneme. Start and end tokens, for phonemes at the start and end of a word, are also counted. After collecting the counts, the values are normalized per row, so all the features for a phoneme sum to 1. Table 3 shows an example of an embedding feature matrix.

2.2.4

Visualization of embedding

Following Meloni *et al.* (2019), to analyze what representation of the phonetic space the embedding encoding learns, we performed hierarchical clustering on embeddings. We computed pairwise euclidean distances between phonemes for the embedding matrix learned from the Dutch portion of the NorthEuraLex dataset and for the phonetic feature matrix from Brown *et al.* (2008). Hierarchical clustering was performed on the distance matrices using neighbour joining (Saitou and Nei 1987). Figure 2 shows the results.

The figure shows that the embedding method groups most vowels together, but also adds some consonants to this group, and places the vowel /ə/ in another group. When looking at the groupings in the embedding encoding, it seems the embedding encoding mainly represents phonemes by their place of articulation rather than by their manner of articulation. /p/ and /m/ are grouped together, both bilabial, but one is a stop or fricative, the other a nasal. /l/ and /r/ are grouped, which are both (apico-)alveolar, but one is an approximant, the other a trill. Groupings on manner, like the stops /t/ and /d/ in the phonetic encoding, are less visible in the embedding encoding.

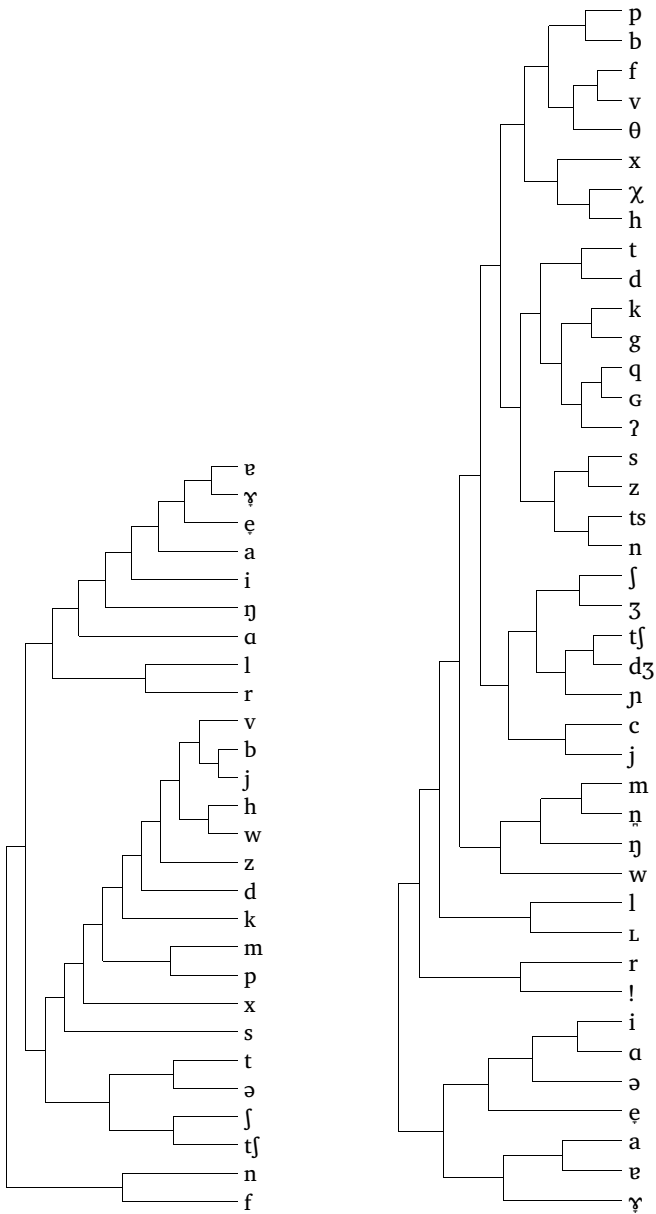


Figure 2:
Hierarchical clusterings,
using euclidean distance
and neighbour joining,
of NorthEuraLex learned
phoneme embedding
for Dutch and the phonetic
feature matrix from Brown
et al. (2008)

(a) Clustering of Dutch
learned embeddings
from NorthEuraLex

(b) Clustering of phonetic
feature matrix from
Brown *et al.* (2008)

Looking at the data, it becomes visible why close phonemes like /d/ and /t/ have more remote embeddings. /d/ occurs in 206 word forms, /t/ in 402 word forms. The position in the words of the two phonemes is quite different. The most frequent position for /d/ (30%) is the first position, in words like: /dʏktər/ “doctor”, /dɛk/ “roof” and /dɛkə/ “blanket”. For /t/, the most frequent position is the 4th position (34%), in words like: /ɛfstɛnt/ “distance”, /lɛftait/ “age” and /hʏxtə/ “height”. This different position of the phonemes in word forms in the corpus can lead to representations in embedding encoding which are not close to each other.

2.2.5

Evaluation of input encodings

With the simple one-hot encoding, the time-tested phonetic feature encoding and the novel embedding encoding, we now have three different ways to represent linguistic items. How well suited are each of these encoding styles for the task of word prediction? We evaluate the three input encodings in combination with two machine learning models (that will be introduced in Section 3): the encoder-decoder and the structured perceptron. Table 4 shows the average word prediction distance over two language families, for the different parameter settings, on the test set. Although the differences in word prediction distance are small, the embedding encoding tends to work best in most test cases. For the Germanic language family, one-hot encoding works slightly better than embedding encoding, but the difference is minimal.

Table 4:
Evaluation of different data encodings, on two models, by word prediction distance (edit distance between prediction and target) for two language families: Slavic and Germanic. The distance is the mean of the distance of all language pairs in the family. Lower distance means better prediction

Method	Language family		
	Slavic	Germanic	
Model	Input encoding		
Enc-dec	One-hot	0.5582	0.5721
Enc-dec	Phonetic	0.5767	0.5853
Enc-dec	Embedding	0.5579	0.5710
Struct perc	One-hot	0.3436	0.4374
Struct perc	Phonetic	0.3465	0.4497
Struct perc	Embedding	0.3423	0.4375

Our second question is: which machine learning models are suitable to perform the task of word prediction? We need a model that can convert sequential input (the source word) to sequential output (the target word). We would like to apply an algorithm that can model the sequential dependencies between consecutive phonemes in a word. We evaluate two of these sequence models, which are both prototypical for a larger class of models: a simple *structured perceptron* (probabilistic sequence model) and a more complex *RNN encoder-decoder* (deep neural network). Furthermore, we will introduce two baseline models, to compare performance.

Structured perceptron

We will look at the *structured perceptron* (Collins 2002; Daume and Marcu 2006), an example of a probabilistic sequence model, which among others has been applied to part-of-speech tagging. The structured perceptron is an extension of a *perceptron* (one-layer neural network) (Rosenblatt 1958) for performing sequential tasks. In this paper, we will evaluate this as one of the models to predict words between languages.

The structured perceptron algorithm is run for I iterations. At every iteration, all N data points are processed. For every input sequence (word, in this case) x , a sequence \hat{y} is predicted, based on the current model parameters \mathbf{w} :

$$(1) \quad \hat{y} = \operatorname{argmax}_{u \in Y} \mathbf{w}^T \phi(x, u)$$

By the *argmax*, the feature function ϕ has to be evaluated for all possible output sequences $u \in Y$; the value which gives the highest output is used as prediction \hat{y} . This *argmax* is computationally expensive, but the Viterbi algorithm (Viterbi 1967) can be run to efficiently estimate the best value \hat{y} .

If the predicted sequence \hat{y} is different from the target sequence y' , the weights are updated using the difference between the feature func-

tion applied to the target and the feature function applied to the predicted value:

$$(2) \quad \mathbf{w} \leftarrow \mathbf{w} + \phi(x, y') - \phi(x, \hat{y})$$

After I iterations, the weights \mathbf{w} of the last iteration are returned. In practice, the *averaged structured perceptron* is used, which outputs an average of the weights over all updates.

We use the implementation from the seqlearn library.⁴ In the experiments, the structured perceptron algorithm is run for 100 iterations of parameter training.

Ciobanu and Dinu (2018) use a Conditional Random Field (Lafferty *et al.* 2001), a structured prediction technique related to the structured perceptron. With this model, predictions between different Romance languages and Latin are made. These pairwise predictions are then ensembled, to arrive at a protoform for Latin.

3.2

RNN encoder-decoder

Deep neural networks have shown recent success in multiple tasks in NLP, like machine translation, using different model architectures, such as the encoder-decoder (Sutskever *et al.* 2014; Cho *et al.* 2014), attention-based models (Bahdanau *et al.* 2014) and transformers (Vaswani *et al.* 2017; Devlin *et al.* 2019). Meloni *et al.* (2019) use an encoder-decoder structure, but add an attention layer, which allows for focusing on segments of the input word that are useful for predicting the target word. The authors would like to predict a Latin word form from a number of contemporary Romance languages. In order to do this, the encoder accepts multiple inputs, one for every language. Fourier and Sagot (2020a) use a multiway encoder-decoder with attention. In this architecture, there is one model for all language pairs, with a separate encoder per source language and a separate decoder per target language. Cathcart and Wandl (2020) apply an encoder-decoder with a specific type of attention, 0th order hard monotonic attention (Wu and Cotterell 2019). The task is to predict contemporary word forms from proto-Slavic word forms. This is done using one

⁴<https://github.com/larsmans/seqlearn>

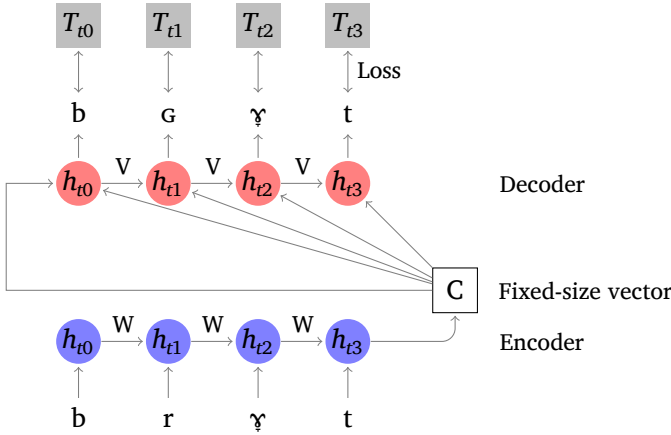


Figure 3:
Structure of our RNN
encoder-decoder model

encoder-decoder model for all languages, where every input word is paired with a language-specific embedding. The language-specific embedding is a straight-through embedding (Bengio *et al.* 2013; Courbariaux *et al.* 2016), which has a discrete representation at prediction time, but also a continuous representation to calculate loss. This enables the interpretation of the neural network as a latent variable model.

In this paper, we consider a relatively simple model, a recurrent neural network (RNN) in encoder-decoder structure, as representative for the class of deep neural networks. A RNN takes a sequence as input and produces a sequence. RNNs are good at handling sequential information because the output of a recurrent node depends on both the input at the current time step (the phoneme at the current position) and on the values of the previous recurrent node, carrying a representation of previous phonemes in the word. An encoder-decoder model consists of two RNNs, see Figure 3. This architecture enables the use of different source and target lengths and outputs a phoneme based on the whole input string.

In our approach, we use a vanilla RNN encoder-decoder, without attention. We evaluated different architectures and parameter settings in preliminary experiments, and picked the best performing (Section 4.1). As recurrent network nodes, we use Gated Recurrent Units (GRU) (Cho *et al.* 2014), which are capable of capturing long-distance dependencies. The GRU is an adaptation of the Long Short

Term Memory (LSTM) unit (Hochreiter and Schmidhuber 1997). Both the encoder and decoder consist of 400 hidden units. The output from the last time step of the encoder is used, a fixed-size (independent of input length) vector. We apply a bidirectional encoder architecture, running a forward-direction and a backward-direction encoder on the input, and combining the output using a dense layer, to reduce it to the dimensionality of the output of a single encoder. The fixed-size vector, which contains information of the forward and backward pass, is then fed to the decoder at every time step.

Because we use one-hot output encoding, predicting a phoneme corresponds to single-label classification: only one element of the vector can be 1. Therefore, the output layer of the network is a *softmax* layer, which outputs a probability distribution over the possible one-hot positions, corresponding to phonemes. The network outputs are compared to the target values using a *categorical cross-entropy* loss function, which is known to work together well with *softmax* output. We add an *L2 regularization term* to the loss function, which penalizes large weight values, to prevent overfitting on the training data.

To give an impression of the degrees of freedom the network has when learning correspondences, we give an estimation of the number of weights in the network. The weights consist of the weights in the encoders, the decoder, the dense encoder concatenation layer and the dense output layer. The number of weights is dependent on the number of features in the encoding chosen for the input and target language, and on the maximum length of words in the languages. For the language pair Dutch-German, with embedding input encoding, there is a total of 2.4 million weights.

The weights of the network are initialized using *Xavier initialization* (Glorot and Bengio 2010). With the right initialization, the network can be trained faster because the incoming data fits better to the activation functions of the layers. We apply dropout, the random disabling of network nodes to prevent overfitting to training data; the dropout factor is 0.1. Data is supplied in batches, the default batch size is 10. The applied optimization algorithm is *Adagrad* (Duchi *et al.* 2011): this is an algorithm to update the weights with the gradient of the loss, using an adaptive learning rate. The initial learning rate is 0.01. The threshold for gradient clipping is set to 100. In the experiments, the default number of training epochs, the number of times

the training set is run through, is 15. The network was implemented using the Lasagne neural network library (Dieleman *et al.* 2015).

Baseline models

3.3

The prediction results are compared to two baselines, for which the distance between target and baseline is calculated. The first baseline is the trivial *source prediction* baseline, predicting exactly the source word.

The second baseline is based on *Pointwise Mutual Information (PMI)* (Church and Hanks 1990; Wieling *et al.* 2009; Jäger 2014; Jäger *et al.* 2017). PMI similarity between words gives high similarity to words with phonemes that are often aligned to each other in the languages. In general, PMI gives higher similarity to words which are cognate, which are predictable through regular sound correspondences (Jäger and Sofroniev 2016). Following the approach in Jäger *et al.* (2017), PMI scores between phonemes are calculated by aligning all words for a language pair in the training set with each other, using Needleman-Wunsch alignment (NW) (Needleman and Wunsch 1970). We perform 50 iterations of NW alignment. At every NW iteration, the weights that determine the match between phonemes, are determined by the PMI scores of the previous iteration. At prediction time, the alignment of the last training iteration is used. For every source phoneme, the target phoneme with the highest probability of being aligned to the source phoneme is predicted. The internal table of PMI scores is essentially a table of sound correspondences the model learns.

Evaluation of machine learning models

3.4

Table 5 shows the results for the two machine learning models, and two baseline models, evaluated on the test set. It can be observed that the structured perceptron, in spite of its simpler structure, performs better than the encoder-decoder. For the Germanic language family, the structured perceptron also performs better than the PMI-based baseline model. For the Slavic language model, the PMI-based baseline works better. This difference may be explained by the fact

Table 5:
Evaluation of machine learning models and baseline models, with one-hot input data, evaluated on word prediction distance (edit distance between prediction and target) for different test conditions, for two language families: Slavic and Germanic. The distance is the mean of the distance of all language pairs in the family. Lower distance means better prediction

Method	Language family	
	Slavic	Germanic
Encoder-decoder	0.5582	0.5721
Structured perceptron	0.3436	0.4374
Source prediction baseline	0.3714	0.4933
PMI-based baseline	0.3249	0.4520

that the Slavic languages in the dataset are more closely related and therefore easier for the baseline model to predict correctly, as smaller changes have to be made to the source words.

4

IMPLEMENTATION DETAILS

Having discussed the different choices for dataset, data representation and machine learning model for word prediction, we will discuss how useful word prediction is in algorithms for computational historical linguistics. We will first, however, briefly describe some of the implementation details needed to get the word prediction models to produce optimal results.

4.1

Parameter optimization

In preliminary experiments, we tested the different models with a range of different parameters and evaluated on a development set. The parameter settings with the highest performance on the development set, were then used for the experiments reported in this paper, in the previous two sections. For the experiments on data encoding and machine learning models in the two previous sections, we used the test set because we regard these as model evaluation, not as parameter tuning. For the experiments in the next sections, on applications, we also use the test set.

Training

4.2

Training is performed on the full training set per language pair, which consists of both cognate and non-cognate words. It would be easier for the model to learn sound correspondences if it would only receive cognate training examples. However, we want to develop a model that can be applied to problems where no cognate judgments are available.

Data preparation

4.3

For the structured perceptron, the input and output word must have the same length, but the lengths of words in a language pair may differ. For this model, the input and output word are matched in terms of length, by padding the shortest word at the end with dummy symbols (.). For the encoder-decoder, all words in a language must have the same fixed length, to fit the fixed shape of the layers, but the input language may have a different fixed length than the output language. To prepare the data for this model, maximum lengths per language are calculated, and words are padded with dummy symbols (.) at the end to match the maximum length.

For the target data for the encoder-decoder model, one-hot encoding is used regardless of the input encoding. This means that target words are encoded in one-hot encoding and the algorithm will output predictions in one-hot encoding. One-hot output encoding facilitates convenient decoding of the predictions. Other output encodings did not show good results in preliminary experiments. As a target for the structured perceptron model, unencoded data is supplied, since this is a format suitable for the used implementation of the algorithm.

The training data is standardized in order to fit it better to the activation functions of the neural network nodes. For the training data, the mean and standard deviation per feature are calculated over the whole training set for this language pair. The mean is subtracted from the data and the resulting value is divided by the standard deviation. After standardization, per feature, the standardized training data has a mean of 0 and a standard deviation of 1. The test data is standardized using the mean and standard deviation of the training data. This transfer of knowledge can be regarded as being part of the training procedure.

We evaluate the models by comparing the predictions and targets on the test section of the dataset. We only evaluate on cognate pairs of words. If words are not genetically related, the algorithm will not be able to predict this word via regular sound correspondences. Cognate judgments from the IELex dataset are used.⁵ For words in NorthEuraLex for which no IELex cognate judgments are available, LexStat (List 2012) automatic cognate judgments are generated (threshold 0.6).

Languages which are not closely related do not share many cognates. Because we only evaluate on cognate words, the test set for those language pairs will become too small. To alleviate this problem, we evaluate only on groups of more closely related languages. In these groups, every language in the group shares at least n cognates with all other languages. We determine these groups, by generating a graph of all languages where two languages are connected if and only if the number of shared cognates exceeds the threshold n . Then, we determine the *maximal cliques* in this graph: groups of nodes where all nodes are connected to each other and it is not possible to add another node that is connected to all existing nodes. These *maximal cliques* correspond to our definition of language groups which share n cognates. The largest cliques were the Slavic (Czech, Bulgarian, Russian, Belarusian, Ukrainian, Polish, Slovak, Slovenian, Croatian) and Germanic (Swedish, Icelandic, English, Dutch, German, Danish, Norwegian) subfamilies, which we use for our experiments. The distance metric used between target and prediction is *normalized edit distance: Levenshtein distance* (Levenshtein 1966) (or: edit distance) divided by the length of the longest sequence. An average of this distance metric, over all words in the test set, is used as the distance between two languages.

We use the prediction distances for two purposes: to determine the distance of languages to each other and to determine the general accuracy of a model. If a certain model has a lower prediction distance

⁵An intersection, which applied the IELex cognate judgments to the NorthEuraLex dataset, was supplied by Gerhard Jäger.

over all language pairs than another model, we consider it to be more accurate.

Qualitative evaluation of word prediction

4.5

Using the best parameters determined in the preceding sections, it is worth looking at some qualitative examples of the predictions the algorithms make. Table 6 shows the output of the word prediction algo-

Input	Target	Prediction	Distance
starktə	ʃtarkə	ʃtarkən	0.14
krais	kɛʏits	kɛɪʃ	0.40
brar	bɔadə	bɔar	0.40
mɔrxə	mɔrgən	mɔrgə	0.17
səmə	tsazəmən	ʃəmə	0.57
varxən	fargəən	fargən	0.14
klimə	klaten	klimə	0.67
bindəl	bindəl	bəndəl	0.17
linkər	liŋkə	liŋkən	0.33
wɔnə	vɔnən	van	0.60
sxəlt	ʃalt	ʃɔəlt	0.40
brendə	bɔanən	bɔəndə	0.50
ski	ski	ʃən	1.00
zikzain	kɛŋkzəin	ziʃzəin	0.56
dreïn	dɛɛən	dɛəin	0.40
rɛxə	ɛɛgən	ɛɛgə	0.20
sidərə	tsitən	ʃidəgə	0.83
halft	halftə	halft	0.17
ɔvarwinə	zigən	ɔfarviən	0.75
fɛrtəx	firtsɪʃ	fartsən	0.50
tɛkə	tsɛɪʃən	tsɛən	0.50
nɛkt	nɛkt	nɛit	0.25
hɔŋɛrix	həŋɛɪʃ	həŋɛɪʃ	0.14
wraivə	ɛəibən	vɛəibə	0.33
zɔndə	zində	zɔndə	0.20
zixvarzəmələ	ziʃfarzəməlɪn	ziʃfarzəmələ	0.08
hɔŋɛr	həŋɛ	həŋən	0.40
zɔmər	zɔmɛ	zɔmən	0.40
hert	harts	hert	0.50
bədrixə	bətɛigən	bədtɛigə	0.25

Table 6:
Word prediction output for a structured perceptron (embedding encoding) on language pair Dutch-German. *Prediction* is the German word predicted by the model when *Input* is given as Dutch input. The edit distance between the prediction and the *target* German word, which is not seen by the model, is calculated. Lower distance is better performance

rithm for a structured perceptron model on the language pair Dutch-German. For every word, a prediction distance is calculated. This distance per word will be used for cognate detection in Subsection 5.3. From these word distances, a mean distance per language pair is calculated. This will be used for the application of phylogenetic tree reconstruction in Subsection 5.1. When again taking the mean of the scores of all language pairs in a family, one could get a score which represents the performance of a model on a language family. These distances were used in the previous sections to evaluate different parameter settings.

5

APPLICATIONS

After we looked at appropriate data and a machine learning model for word prediction, it is now time to discuss a number of applications of word prediction in historical linguistics: *phylogenetic tree reconstruction*, *sound correspondence identification* and *cognate detection*. The applications use the outcomes of word prediction as a basis. After describing each application, we will evaluate the results.

5.1

Phylogenetic tree reconstruction

We regard the prediction score between language pairs as a measure of ancestral relatedness and use these scores to reconstruct a phylogenetic tree (see Section 6.3 for a further discussion). We perform hierarchical clustering on the matrix of edit distances for all language pairs, using the *UPGMA* (Sokal and Michener 1958) and *neighbour joining* (Saitou and Nei 1987) algorithms, implemented in the `LingPy` library (List *et al.* 2019). The generated trees are then compared to reference trees from Glottolog (Hammarström *et al.* 2020), based on current insights in historical linguistics. Evaluation is performed using Generalized Quartet Distance (Pompei *et al.* 2011), a generalization of Quartet Distance (Bryant *et al.* 2000) to non-binary trees. We apply the algorithm as implemented in the `QDist` software package (Mailund and

Method	Clustering	
	UPGMA	Neighbour joining
Struct perc (embedding enc)	0.047619	0.047619
Source prediction baseline	0.269841	0.047619
PMI-based baseline	0.047619	0.047619

Table 7: Generalized Quartet distance between trees of the Slavic language family, inferred from word prediction results using the structured perceptron model, and the Glottolog reference tree. Lower is better: a generated tree equal to the reference tree will have a theoretical distance of 0. In this case, the lower bound is 0.047619 because the generated binary trees will never precisely match the multiple-branching reference tree

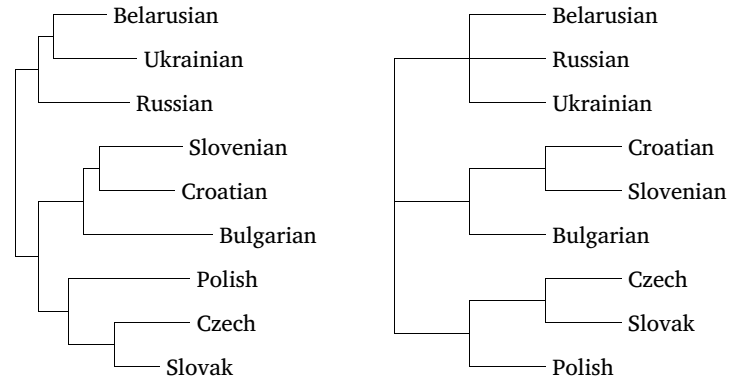
Pedersen 2004). Trees were visualized using the ete3 library (Huerta-Cepas *et al.* 2016).

This is the first approach to infer phylogenetic trees from prediction distance. Cathcart and Wandl (2020) infer phylogenetic trees, based on the language-specific embedding vectors, which are paired with the input and trained with the model. Distances between embedding vectors are calculated using cosine distance, and the resulting distance matrix is clustered using neighbour joining.

Table 7 shows the generalized Quartet distance between the generated trees, for different conditions, and a Glottolog reference tree. In the table, one could see that the structured perceptron model consistently creates valid trees. The baseline models, especially the PMI model, also create valid trees. The performance differences between models are smaller than the differences for the word prediction task. This is not very surprising, given that phylogenetic tree reconstruction is an easier task than word prediction: there are fewer possible branchings in a tree, than possible combinations of phonemes in a word. Even a model with lower performance on word prediction can generate a relatively good tree.

Figure 4 graphically shows the trees inferred from word prediction using the structured perceptron. The Glottolog reference tree is added for comparison. The perceptron tree receives the lowest possible distance to the reference tree of 0.047619: the generated binary trees will never precisely match the multiple-branching reference tree. This is also the reason why no generated tree reaches the Quartet distance of 0 in Table 7.

Figure 4:
Phylogenetic trees for the Slavic language family, using structured perceptron prediction, and the Glottolog reference tree. Quartet distance between trees: 0.047619



(a) Structured perceptron, embedding encoding, NJ clustering

(b) Glottolog

5.2

Sound correspondence identification

To be able to make predictions, the word prediction model has to learn the probabilities of phonemes changing into other phonemes, given a certain context. We would like to extract these correspondences from the model. It is challenging to identify specific neural network nodes that fire when a certain sound correspondence is applied. Instead, we estimate the *internal* sound correspondences that the network learned, by looking at the *output*: the substitutions made between the source word and the prediction. Pairs of source and prediction words are aligned using the Needleman-Wunsch algorithm. Then, the pairs of substituted phonemes between these source-prediction alignments can be counted. These can be compared to the counts of substituted phonemes between source and target. In other prediction approaches, sound correspondences are not directly determined. Instead of listing sound correspondences, Meloni *et al.* (2019) and Cathcart and Wandl (2020) make an analysis of the errors that the algorithm makes per phonological phenomenon, evaluated on both real and synthetic data.

We identified sound correspondences between Dutch and German, two closely related Germanic languages. Source-prediction substitutions were extracted using a structured perceptron model, run for

Substitution	Source-prediction frequency	Source-target frequency
r	g	37
s	ʃ	27
x	g	23
v	f	16
a	ɐ	16
w	v	11
–	n	10
t	ts	9
ʎ	ɑ	8
r	n	6
p	f	6
ɛ	a	5
x	ʃ	5
ə	–	4
a	ə	4
a	i	3
i	ə	3
n	–	3
t	–	3
v	b	3

Table 8: Substitutions between aligned source-prediction pairs and substitutions between aligned source-target pairs for Dutch-German word prediction, using a structured perceptron model and embedding encoding. The list is ordered on frequency of source-prediction substitutions, the 20 most frequent entries are shown

the default 100 iterations (Section 3.1). Table 8 shows the most frequent sound substitutions for source-prediction and source-target. It can be observed that the most frequent substitutions between source and prediction are also frequent between source and target. This implies that the model learned meaningful sound correspondences.

Cognate detection

5.3

Cognate detection is the detection of word forms in different languages (usually per concept), which derive from the same ancestral word. In order to perform cognate detection based on word prediction, we

cluster words for the same concept in different languages based on the prediction distances per word.

First, word prediction is performed for all language pairs which we want to evaluate. In the normal word prediction workflow (Section 4.4), predictions are made only on word pairs which are deemed cognates by existing judgments. When performing cognate detection, the whole point is to make these judgments, so we perform word prediction on the full test set: cognates and non-cognates. We take into account concepts for which word forms occur in all languages; this vastly reduces the number of concepts. For every concept, we create a distance matrix between the word forms in all languages, based on the prediction distance *per word*. Next, we run a flat clustering algorithm on this distance matrix. Applied clustering algorithms are *flat UPGMA* (Sokal and Michener 1958), *link clustering* (Ahn *et al.* 2010) and *MCL* (van Dongen 2000), implemented in the LingPy library. Preliminary experiments on the development set show that a threshold of $\theta = 0.7$ gives best results for MCL and link clustering, and $\theta = 0.8$ gives best results for flat UPGMA.

Conceptually, the performed cognate detection operation is the same as the phylogenetic tree reconstruction operation, but now we cluster per word, instead of per language, and we perform a flat clustering instead of an hierarchical clustering.

For evaluation, we use cognate judgments from IElex (Dunn 2012). Evaluation is performed using the B-Cubed F measure (Bagga and Baldwin 1998; Amigó *et al.* 2009), implemented in the bcubed library.⁶ We perform cognate detection for the Slavic and Germanic language families, by clustering words based on word prediction distances. We evaluate performance for the structured perceptron model, compared to the source prediction baseline. During cognate detection, contrary to the default setting, prediction is performed on both cognates and non-cognates. We apply three clustering algorithms: *MCL* ($\theta = 0.7$), *Link clustering* ($\theta = 0.7$) and *Flat UPGMA* ($\theta = 0.8$).

Table 9 shows B-Cubed F scores for cognate detection on the Slavic and Germanic language families. For the Germanic language family, the structured perceptron, using MCL clustering, performs best. For the Slavic language family, the source prediction baseline

⁶<https://github.com/hhromic/python-bcubed>

Model	Cluster algorithm	Language family	
		Slavic	Germanic
Structured perceptron (one-hot)	MCL	0.877458	0.932077
Structured perceptron (one-hot)	LC	0.915680	0.905044
Structured perceptron (one-hot)	fUPGMA	0.919739	0.889788
Source prediction	MCL	0.920806	0.851781
Source prediction	LC	0.926061	0.875475
Source prediction	fUPGMA	0.929840	0.878705

Table 9: B-Cubed F scores for cognate detection on the Slavic (27 concepts) and Germanic (29 concepts) language families. MCL = MCL clustering ($\theta = 0.7$), LC = Link Clustering ($\theta = 0.7$), fUPGMA = Flat UPGMA ($\theta = 0.8$). Higher F score means better correspondence between computed and real clustering

model slightly outperforms the structured perceptron. It must be noted that the sample of shared concepts in a language family is small: this makes results less stable.

DISCUSSION

6

Contribution

6.1

In this paper, we evaluated under which conditions the machine learning paradigm, successful in many computing tasks, can be useful in historical linguistics. We proposed the task of *word prediction*: by training a machine learning model on pairs of words in two languages, it learns the sound correspondences between the two languages and should be able to predict unseen words. We regard this as a method that stays close to the central aspects of the traditional comparative method in historical linguistics and is therefore a good candidate for reliable reconstruction of language ancestry. Multiple factors which could lead to an effective use of prediction methods in historical linguistics were evaluated: the choice of machine learning model and encoding of the input data. We evaluated existing models of word prediction (Ciobanu and Dinu 2018; Meloni *et al.* 2019; Cathcart and Wandl 2020; Fourier

and Sagot 2020a) and came up with our own model, which enables applications on several tasks in historical linguistics. In this paper, we have proposed new approaches for phylogenetic tree reconstruction and cognate detection, based on word prediction error. We evaluated two sequential neural network models, a RNN encoder-decoder and a structured perceptron. To find an appropriate data representation, we evaluated embedding encoding, inspired by word embeddings in natural language processing, and compared its performance to existing one-hot and phonetic encodings. Our results suggest that a simple structured perceptron performs better than a RNN encoder-decoder, and embedding encoding performs slightly better than existing encodings on the prediction task. It should also be noted that one of our baselines, the PMI-based prediction model, performs relatively well, probably because this simple method does have an internal representation of sound correspondences. More research is needed to find the exact model architectures, parameter settings and data encodings to obtain optimal performance in the word prediction tasks. Note that the goal of our current paper is merely exploratory: we explore the conditions under which the prediction paradigm can be used in historical linguistics, and what possible applications of prediction could be in historical linguistics.

6.2

Related work

We will now discuss two types of related work. Firstly, we will look at other approaches in computational historical linguistics which try to capture a genotypic relationship. Secondly, we will look at other approaches which use the concept of *word prediction*, in multiple contexts.

6.2.1

Genotypic methods in computational historical linguistics

Many approaches in computational historical linguistics try to capture a genetic signal, by staying close to one or more steps of the comparative method, where sound correspondences are a central notion. Hruschka *et al.* (2015) create a phylogeny of languages using Bayesian MCMC, while at the same time giving a probabilistic description of

regular sound correspondences. By explicitly modelling the sound correspondences, this approach stays close to one of the main principles of the comparative method. Bouchard-Côté *et al.* (2013) directly compare phonetic strings of words, in order to reconstruct the protoforms of words and perform cognate detection. Probabilistic string transducers model the sound changes, taking into account context. A tree is postulated, and in an iterative process, candidate protoforms are generated. Parameters are estimated using Expectation Maximization. Sound correspondences and protoforms, central notions in the comparative method, are explicitly modelled.

In cognate detection, some approaches depart from comparing phonetic forms, but then follow a genotypic path, by extracting sound correspondences. List (2012) places phonetic strings of words into sound classes. Then, a matrix of language-pair dependent scores for sound correspondences is extracted. Based on this matrix, distances are assigned to cognate candidates. Finally, they are clustered into cognate classes. The matrix that is internally kept, is essentially a matrix of sound correspondences.

Different approaches are applied to follow the comparative method in phylogenetic tree reconstruction. Jäger (2015) applies a distance-based clustering algorithm for tree reconstruction. String similarities between alignments of words are directly used as distances between the languages. Although direct string comparison looks like a phenotypic step in this method, common ancestry is captured by explicitly removing words which show chance resemblances and borrowings, using a statistical test (Cronbach's alpha). Jäger (2018) introduces *soundclass-concept* characters to encode word forms as input for character-based phylogenetic models. In this approach, a word is represented by the presence or absence of (classes of) phonemes, leading to a representation of sound changes.

Approaches similar to word prediction have been applied before in the natural language processing community, but with differences in implementation and goal to our approach. An early example is Mulloni (2007), who used Support Vector Machine classifiers to predict words, with the goal of improving bilingual terminology lists and machine

translation algorithms. Beinborn *et al.* (2013) and Ciobanu (2016) perform word prediction, using methods from statistical machine translation, to assess the learnability of words for second language learners.

Some recent prediction approaches from NLP have the goal, like our approach, to reconstruct language ancestry. Ciobanu and Dinu (2018) predict from several Romance languages to Latin using Conditional Random Fields, and perform preliminary experiments using RNNs. Results for the different language pairs are then combined using an ensemble system to arrive at Latin protoforms. Ciobanu *et al.* (2020) and Ciobanu and Dinu (2020) report on variations of the same method, for prediction to modern languages. Meloni *et al.* (2019) use encoder-decoder neural networks with attention to predict Latin forms, from word lists from multiple Romance languages simultaneously as input. An analysis is performed of the structures the network learned, by evaluating the model on synthetic input words. Cathcart and Wandl (2020) predict words, in phonetic form, from proto-Slavic protoforms to contemporary Slavic languages, using an encoder-decoder with attention. The authors perform an elaborate error analysis, and use the trained embeddings of their model to reconstruct a phylogenetic tree of Slavic languages. Fourrier and Sagot (2020a) and Fourrier (2020) predict words back and forth between contemporary and proto-languages, and between contemporary languages, using artificial and realistic data, applying a model from statistical machine translation and a neural multiway encoder-decoder.

Although the applied methods differ, what the preceding approaches have in common is that their input consists solely of cognates. In our approach, the algorithm can be trained on data which is not labelled for cognacy, avoiding the need for manual cognate judgments. Moreover, in our approach, prediction results serve as a starting point for performing a number of diverse tasks in historical linguistics, such as phylogenetic tree reconstruction and cognate detection.

Recently, the prediction paradigm has gained ground in the computational historical linguistics community as well. List (2019a) uses a network analysis algorithm to predict word forms. This method is evaluated by predicting forms for unexplored languages, which are then attested by performing linguistic fieldwork (Bodt and List 2019, 2020).

Is the prediction error (normalized edit distance) between the predicted word and the target word a good measure for the distance between languages? The intuition behind using the prediction error as language distance is that two languages which are well-predictable through regular sound correspondences, must be closely related. There are however a few issues involved.

Firstly, assigning a distance of 0 for word forms which are fully predictable using regular sound correspondences, is somewhat problematic. One could argue that languages which differ only through regular sound correspondences should also receive a non-zero distance because they are not identical. For the specific case of reconstructing proto-languages, List (2019b) proposes that two reconstruction systems for a proto-language, that produce protoforms only differing by regular correspondences, can be seen as *structurally identical*. The rationale behind the concept of structural identity is that reconstructions of proto-languages are to some extent abstractions, in which arbitrary symbols could be used in protoforms. However, contemporary languages only differing through regular sound correspondences could not be called structurally identical, as these languages are not abstractions. Ideally, a model of language ancestry would give multiple distances: one distance based on the number of mutations made using regular sound correspondences, and one distance based on the number of irregular mutations made (including non-cognate words). Hruschka *et al.* (2015) created a model which explicitly models the distinction between regular and irregular sound changes, when creating a phylogenetic family tree. This model does not, however, rely on prediction nor prediction distances.

Another issue is that prediction error is not a very informative distance for languages which have non-cognate word pairs for many concepts. For non-cognate word pairs, the model cannot apply any learned sound correspondences to predict the word. The word will in many cases be completely incorrectly predicted, with a prediction error (normalized edit distance) towards 1. This does not inform us about the linguistic distance between these words. As most non-cognate word pairs will receive an error towards 1, when averaging over all con-

cepts for a language, this will give a measure of the proportion of non-cognates between two languages.

The final issue that needs to be discussed is that, when using the prediction error as language distance, the prediction error also signifies the model performance. When using a better-performing model, the distances between some or all languages may suddenly be smaller. Ideally, one would want a metric of language distance which is independent of the performance of the underlying model. A distance metric that discounts for model complexity could possibly draw upon ideas from the Minimum Description Length principle (MDL) (Rissanen 1978; Grunwald 2004). In the MDL framework, the best model to describe a dataset is the simplest model that is accurately able to compress the data by finding regularities. This corresponds to the model with the lowest description length. The description length is the sum of the length it takes to describe the model (model complexity) and the length it takes to describe the data with the help of the model (prediction error). In our case, when using description length as a distance metric, the low prediction error of a complex model will be discounted by adding the model complexity to the distance. MDL has been used before as a distance metric between languages to perform phylogenetic reconstruction (Wettig *et al.* 2011; Fischer *et al.* 2018).

All in all, prediction error is not a perfect measure for language distance. However, it is a reasonable approximation for our purposes, which is straightforward to obtain from a prediction model.

6.4

Historical linguistics as latent variable model

Taking a step back, the problem of inferring the phylogeny of languages from present-day language data can be viewed as a *latent variable model*. A latent variable model is a model where latent variables η , whose value cannot be observed, are connected to observed variables \mathbf{y} . In these models, one could infer the value of a hidden variable from a certain observed variable. As a genotypic method, word prediction is one instantiation of the latent variable problem of inferring phylogeny, although it has not been explicitly modelled as such a problem in this paper. Cathcart and Wandl (2020) (cf. earlier approaches

Doyle *et al.* (2014); Murawaki (2017)), explicitly model latent variables to describe language history, by using a neural network with discrete straight-through embeddings. A prediction model, where latent variables can be identified, such as phonological processes of change, appears to be a viable direction for the future.

CONCLUSION

7

With this paper, we hope to contribute to future insights about the ancestry of languages. By applying computational methods in historical linguistics, advances have been made in recent years. In this paper, we built further upon this development and proposed a central role for the prediction paradigm from machine learning in historical linguistics. We showed that a simple probabilistic sequence model and embedding encoding of input data can be good implementation choices. We came up with approaches to apply the prediction paradigm to multiple tasks in historical linguistics: phylogenetic tree reconstruction, sound correspondence identification and cognate detection. After validating these techniques on well-studied language families, they can be especially valuable for language families for which ample data is available, but the exact language history remains unclear. We are looking forward to future research on prediction methods in historical linguistics that can further explore good computational models to come to new linguistic insights.

CODE

8

A user-friendly, interactive version of the code, in a Jupyter notebook, can be downloaded from <https://github.com/peterdekker/prediction-histling/>. This code is meant for educational purposes, results may differ slightly from those presented in this paper. For the original code used to generate the results in this paper, see <https://bitbucket.org/pdekker/wordprediction/>.

ACKNOWLEDGEMENTS

We would like to thank Gerhard Jäger, Johann-Mattis List, Guus Kroonen and Bart de Boer for their valuable comments. PD was supported by a PhD Fellowship fundamental research (11A2821N) of the Research Foundation – Flanders (FWO) and by funding from the Flemish Government under the *Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen* programme. WZ was funded by the Netherlands Organization for Scientific Research (NWO), through a Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

REFERENCES

- Yong-Yeol AHN, James P. BAGROW, and Sune LEHMANN (2010), Link Communities Reveal Multiscale Complexity in Networks, *Nature*, 466(7307):761–764.
- Enrique AMIGÓ, Julio GONZALO, Javier ARTILES, and Felisa VERDEJO (2009), A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints, *Information Retrieval*, 12(4):461–486.
- Cormac ANDERSON, Tiago TRESOLDI, Thiago CHACON, Anne-Maria FEHN, Mary WALWORTH, Robert FORKEL, and Johann-Mattis LIST (2018), A Cross-Linguistic Database of Phonetic Transcription Systems, *Yearbook of the Poznan Linguistic Meeting*, 4(1):21–53, ISSN 2449-7525, doi:10.2478/yplm-2018-0002.
- Amit BAGGA and Breck BALDWIN (1998), Entity-Based Cross-Document Coreferencing Using the Vector Space Model, in *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pp. 79–85.
- Dzmitry BAHDANAU, Kyunghyun CHO, and Yoshua BENGIO (2014), Neural Machine Translation by Jointly Learning to Align and Translate, *arXiv preprint arXiv:1409.0473*.
- Lisa BEINBORN, Torsten ZESCH, and Iryna GUREVYCH (2013), Cognate Production Using Character-Based Machine Translation, in Ruslan MITKOV and Jong C. PARK, editors, *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 883–891, Nagoya, Japan.
- Yoshua BENGIO, Nicholas LÉONARD, and Aaron COURVILLE (2013), Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation, *arXiv:1308.3432 [cs]*.

- Timotheus A. BODT and Johann-Mattis LIST (2019), Testing the Predictive Strength of the Comparative Method: An Ongoing Experiment on Unattested Words in Western Kho-Bwa Languages, *Papers in Historical Phonology*, 4:22–44, ISSN 2399-6714, doi:10.2218/pihph.4.2019.3037.
- Timotheus A. BODT and Johann-Mattis LIST (2020), The Multiple Benefits of Making Predictions in Linguistics, *Babel: The Language Magazine*, 31(2):8–12, doi:http://dx.doi.org/10.17613/m688-4b90.
- Alexandre BOUCHARD-CÔTÉ, David HALL, Thomas L. GRIFFITHS, and Dan KLEIN (2013), Automated Reconstruction of Ancient Languages Using Probabilistic Models of Sound Change, *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.
- Alexandre BOUCHARD-CÔTÉ, Percy LIANG, Thomas L. GRIFFITHS, and Dan KLEIN (2007), A Probabilistic Approach to Diachronic Phonology, in Jason EISNER, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 887–896.
- Remco BOUCKAERT, Philippe LEMEY, Michael DUNN, Simon J. GREENHILL, Alexander V. ALEKSEYENKO, Alexei J. DRUMMOND, Russell D. GRAY, Marc A. SUCHARD, and Quentin D. ATKINSON (2012), Mapping the Origins and Expansion of the Indo-European Language Family, *Science*, 337(6097):957–960.
- Cecil H. BROWN, Eric W. HOLMAN, Søren WICHMANN, and Viveka VELUPILLAI (2008), Automated Classification of the World’s Languages: A Description of the Method and Preliminary Results, *STUF – Language Typology and Universals*, 61(4):285–308.
- David BRYANT, John TSANG, Paul E. KEARNEY, and Ming LI (2000), Computing the Quartet Distance between Evolutionary Trees, in *Symposium on Discrete Algorithms: Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*, volume 9, pp. 285–286.
- Lyle CAMPBELL (2013), *Historical Linguistics: An Introduction*, MIT Press, second edition.
- Chundra CATHCART and Taraka RAMA (2020), Disentangling Dialects: A Neural Approach to Indo-Aryan Historical Phonology and Subgrouping, in Raquel FERNÁNDEZ and Tal LINZEN, editors, *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 620–630, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.conll-1.50.
- Chundra CATHCART and Florian WANDL (2020), In Search of Isoglosses: Continuous and Discrete Language Embeddings in Slavic Historical Phonology, in Garrett NICOLAI, Kyle GORMAN, and Ryan COTTERELL, editors, *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 233–244, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.sigmorphon-1.28.

Will CHANG, Chundra CATHCART, David HALL, and Andrew GARRETT (2015), Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis, *Language*, 91(1):194–244.

Kyunghyun CHO, Bart VAN MERRIËNBOER, Caglar GULCEHRE, Dzmitry BAHDANAU, Fethi BOUGARES, Holger SCHWENK, and Yoshua BENGIO (2014), Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation, in Alessandro MOSCHITTI, Bo PANG, and Walter DAELEMANS, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Association for Computational Linguistics, Doha, Qatar, doi:10.3115/v1/D14-1179.

Kenneth Ward CHURCH and Patrick HANKS (1990), Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, 16(1):22–29.

Alina Maria CIOBANU (2016), Sequence Labeling for Cognate Production, *Procedia Computer Science*, 96:1391–1399, ISSN 18770509, doi:10.1016/j.procs.2016.08.184.

Alina Maria CIOBANU and Liviu P. DINU (2014), Building a Dataset of Multilingual Cognates for the Romanian Lexicon, in Nicoletta CALZOLARI, Khalid CHOUKRI, Thierry DECLERCK, Hrafn LOFTSSON, Bente MAEGAARD, Joseph MARIANI, Asuncion MORENO, Jan ODIJK, and Stelios PIPERIDIS, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 1038–1043.

Alina Maria CIOBANU and Liviu P. DINU (2018), Ab Initio: Automatic Latin Proto-Word Reconstruction, in Emily M. BENDER, Leon DERCZYNSKI, and Pierre ISABELLE, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1604–1614.

Alina Maria CIOBANU and Liviu P. DINU (2020), Automatic Identification and Production of Related Words for Historical Linguistics, *Computational Linguistics*, 45(4):667–704, ISSN 0891-2017, 1530-9312, doi:10.1162/coli_a_00361.

Alina Maria CIOBANU, Liviu P. DINU, and Laurentiu ZOICAS (2020), Automatic Reconstruction of Missing Romanian Cognates and Unattested Latin Words, in *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 3226–3231.

James CLACKSON (2007), *Indo-European Linguistics: An Introduction*, Cambridge University Press.

Michael COLLINS (2002), Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms, in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, pp. 1–8.

Matthieu COURBARIAUX, Itay HUBARA, Daniel SOUDRY, Ran EL-YANIV, and Yoshua BENGIO (2016), Binarized Neural Networks: Training Deep Neural

Networks with Weights and Activations Constrained to +1 or -1, *arXiv:1602.02830 [cs]*.

Harold Charles DAUME and Daniel MARCU (2006), *Practical Structured Learning Techniques for Natural Language Processing*, University of Southern California.

Peter DEKKER (2018), *Reconstructing Language Ancestry by Performing Word Prediction with Neural Networks*, MSc thesis, University of Amsterdam.

Johannes DELLERT (2018), Combining Information-Weighted Sequence Alignment and Sound Correspondence Models for Improved Cognate Detection, in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3123–3133.

Johannes DELLERT, Thora DANAYKO, Alla MÜNCH, Alina LADYGINA, Armin BUCH, Natalie CLARIUS, Ilja GRIGORJEW, Mohamed BALABEL, Hizniye Isabella BOGA, Zalina BAYSAROVA, Roland MÜHLENBERND, Johannes WAHLE, and Gerhard JÄGER (2019), NorthEuraLex: A Wide-Coverage Lexical Database of Northern Eurasia, *Language Resources and Evaluation*, ISSN 1574-020X, 1574-0218, doi:10.1007/s10579-019-09480-6.

Rick DERKSEN (2007), *Etymological Dictionary of the Slavic Inherited Lexicon*, Brill.

Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE, and Kristina TOUTANOVA (2019), BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, in Jill BURSTEIN, Christy DORAN, and Thamar SOLORIO, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota, doi:10.18653/v1/N19-1423.

Sander DIELEMAN, Jan SCHLÜTER, Colin RAFFEL, Eben OLSON, Søren Kaae SØNDERBY, Daniel NOURI, Daniel MATURANA, Martin THOMA, Eric BATTENBERG, Jack KELLY, Jeffrey De FAUW, Michael HEILMAN, Diogo Moitinho DE ALMEIDA, Brian MCFEE, Hendrik WEIDEMAN, Gábor TAKÁCS, Peter DE RIVAZ, Jon CRALL, Gregory SANDERS, Kashif RASUL, Cong LIU, Geoffrey FRENCH, and Jonas DEGRAVE (2015), Lasagne: First Release., doi:10.5281/zenodo.27878.

Gabriel DOYLE, Klinton BICKNELL, and Roger LEVY (2014), Nonparametric Learning of Phonological Constraints in Optimality Theory, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1094–1103, Association for Computational Linguistics, Baltimore, Maryland, doi:10.3115/v1/P14-1103.

John DUCHI, Elad HAZAN, and Yoram SINGER (2011), Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *Journal of Machine Learning Research*, 12(Jul):2121–2159.

- Michael DUNN (2012), Indo-European Lexical Cognacy Database (IELex).
- John R. FIRTH (1957), A Synopsis of Linguistic Theory, 1930-1955, *Studies in linguistic analysis*.
- Andrea K. FISCHER, Jilles VREEKEN, and Dietrich KLAOW (2018), Beyond Pairwise Similarity: Quantifying and Characterizing Linguistic Similarity between Groups of Languages by MDL, *Computación y Sistemas*, 21(4), ISSN 2007-9737, 1405-5546, doi:10.13053/cys-21-4-2865.
- Robert FORKEL, Johann-Mattis LIST, Simon J. GREENHILL, Christoph RZYMSKI, Sebastian BANK, Michael CYSOUW, Harald HAMMARSTRÖM, Martin HASPELMATH, Gereon A. KAIPING, and Russell D. GRAY (2018), Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics, *Scientific Data*, 5:180205, ISSN 2052-4463, doi:10.1038/sdata.2018.205.
- Clémentine FOURRIER (2020), Évolution phonétique des langues et réseaux de neurones: travaux préliminaires, in Christophe BENZITOUN, Chloé BRAUD, Laurine HUBER, David LANGLOIS, Slim OUNI, and Sylvain POGODALLA, editors, *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*, volume 3: Rencontre des Étudiants Chercheurs en Informatique pour le TAL.
- Clémentine FOURRIER and Benoît SAGOT (2020a), Comparing Statistical and Neural Models for Learning Sound Correspondences, in *LT4HALA 2020: First Workshop on Language Technologies for Historical and Ancient Languages*, Marseille, France.
- Clémentine FOURRIER and Benoît SAGOT (2020b), Methodological Aspects of Developing and Managing an Etymological Lexical Resource: Introducing EtymDB-2.0, in Nicoletta CALZOLARI, Frédéric BÉCHET, Philippe BLACHE, Khalid CHOUKRI, Christopher CIERI, Thierry DECLERCK, Sara GOGGI, Hitoshi ISAHARA, Bente MAEGAARD, Joseph MARIANI, Hélène MAZO, Asuncion MORENO, Jan ODIJK, and Stelios PIPERIDIS, editors, *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 3207–3216, European Language Resources Association, Marseille, France, ISBN 979-10-95546-34-4.
- Xavier GLOROT and Yoshua BENGIO (2010), Understanding the Difficulty of Training Deep Feedforward Neural Networks, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256.
- Russell D. GRAY and Quentin D. ATKINSON (2003), Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin, *Nature*, 426(6965):435–439.
- Simon J. GREENHILL, Chieh-Hsi WU, Xia HUA, Michael DUNN, Stephen C. LEVINSON, and Russell D. GRAY (2017), Evolutionary Dynamics of Language Systems, *Proceedings of the National Academy of Sciences*, 114(42):E8822–E8829.

- Peter GRUNWALD (2004), A Tutorial Introduction to the Minimum Description Length Principle, *arXiv:math/0406077*.
- Harald HAMMARSTRÖM, Robert FORKEL, Martin HASPELMATH, and Sebastian BANK (2020), Glottolog 4.2.1.
- Sepp HOCHREITER and Jürgen SCHMIDHUBER (1997), Long Short-Term Memory, *Neural Computation*, 9(8):1735–1780.
- Daniel J. HRUSCHKA, Simon BRANFORD, Eric D. SMITH, Jon WILKINS, Andrew MEADE, Mark PAGEL, and Tanmoy BHATTACHARYA (2015), Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution, *Current Biology*, 25(1):1–9.
- Jaime HUERTA-CEPAS, François SERRA, and Peer BORK (2016), ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data, *Molecular Biology and Evolution*, 33(6):1635–1638.
- Diana INKPEN, Oana FRUNZA, and Grzegorz KONDRAK (2005), Automatic Identification of Cognates and False Friends in French and English, in *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 251–257.
- Gerhard JÄGER (2014), Phylogenetic Inference from Word Lists Using Weighted Alignment with Empirically Determined Weights, in *Quantifying Language Dynamics*, pp. 155–204, Brill.
- Gerhard JÄGER (2015), Support for Linguistic Macrofamilies from Weighted Sequence Alignment, *Proceedings of the National Academy of Sciences*, 112(41):12752–12757.
- Gerhard JÄGER (2018), Global-Scale Phylogenetic Linguistic Inference from Lexical Resources, *Scientific Data*, 5(1):180189, ISSN 2052-4463, doi:10.1038/sdata.2018.189.
- Gerhard JÄGER (2019), Computational Historical Linguistics, *Theoretical Linguistics*, 45(3–4):151–182, ISSN 0301-4428, 1613-4060, doi:10.1515/tl-2019-0011.
- Gerhard JÄGER and Johann-Mattis LIST (2016), Statistical and Computational Elaborations of the Classical Comparative Method.
- Gerhard JÄGER, Johann-Mattis LIST, and Pavel SOFRONIEV (2017), Using Support Vector Machines and State-of-the-Art Algorithms for Phonetic Alignment to Identify Cognates in Multi-Lingual Wordlists, in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1205–1216, Association for Computational Linguistics, Valencia, Spain, doi:10.18653/v1/E17-1113.
- Gerhard JÄGER and Pavel SOFRONIEV (2016), Automatic Cognate Classification with a Support Vector Machine, in *Proceedings of the 13th Conference on Natural Language Processing*, volume 16.

Yoon KIM, Yacine JERNITE, David SONTAG, and Alexander M. RUSH (2016), Character-Aware Neural Language Models, in *Thirtieth AAAI Conference on Artificial Intelligence*.

John D. LAFFERTY, Andrew MCCALLUM, and Fernando C. N. PEREIRA (2001), Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, pp. 282–289, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ISBN 1-55860-778-1.

Roger LASS (1997), *Historical Linguistics and Language Change*, Cambridge University Press, Cambridge.

Vladimir I. LEVENSHEIN (1966), Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, in *Soviet Physics Doklady*, volume 10, pp. 707–710.

Johann-Mattis LIST (2012), LexStat: Automatic Detection of Cognates in Multilingual Wordlists, in Miriam BUTT, Sheelagh CARPENDALE, Gerald PENN, Jelena PROKIĆ, and Michael CYSOUW, editors, *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pp. 117–125, Association for Computational Linguistics, Avignon, France.

Johann-Mattis LIST (2019a), Automatic Inference of Sound Correspondence Patterns across Multiple Languages, *Computational Linguistics*, 45(1):137–161, ISSN 0891-2017, 1530-9312, doi:10.1162/coli_a_00344.

Johann-Mattis LIST (2019b), Beyond Edit Distances: Comparing Linguistic Reconstruction Systems, *Theoretical Linguistics*, 45(3-4):247–258, ISSN 0301-4428, 1613-4060, doi:10.1515/tl-2019-0016.

Johann-Mattis LIST, Simon GREENHILL, Tiago TRESOLDI, and Robert FORKEL (2019), LingPy. A Python Library for Historical Linguistics, *Jena: Max Planck Institute for the Science of Human History*, doi:<https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy>.

Thomas MAILUND and Christian NS PEDERSEN (2004), QDist – Quartet Distance between Evolutionary Trees, *Bioinformatics*, 20(10):1636–1637.

Carlo MELONI, Shauli RAVFOGEL, and Yoav GOLDBERG (2019), Ab Antiquo: Proto-Language Reconstruction with RNNs, *arXiv:1908.02477 [cs]*.

Tomas MIKOLOV, Ilya SUTSKEVER, Kai CHEN, Greg S. CORRADO, and Jeff DEAN (2013), Distributed Representations of Words and Phrases and Their Compositionality, in *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Andrea MULLONI (2007), Automatic Prediction of Cognate Orthography Using Support Vector Machines, in Chris BIEMANN, Violeta SERETAN, and Ellen RILOFF, editors, *Proceedings of the ACL 2007 Student Research Workshop*, pp. 25–30, Association for Computational Linguistics, Prague, Czech Republic.

Yugo MURAWAKI (2017), Diachrony-Aware Induction of Binary Latent Representations from Typological Features, in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, volume 1: Long papers, pp. 451–461.

Saul B. NEEDLEMAN and Christan D. WUNSCH (1970), A Gene Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *Journal of Molecular Biology*, 48:443–453.

Jeffrey PENNINGTON, Richard SOCHER, and Christopher MANNING (2014), Glove: Global Vectors for Word Representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Simone POMPEI, Vittorio LORETO, and Francesca TRIA (2011), On the Accuracy of Language Trees, *PLoS One*, 6(6):e20109.

Taraka RAMA (2016), Siamese Convolutional Networks for Cognate Identification, in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*.

Taraka RAMA and Johann-Mattis LIST (2019), An Automated Framework for Fast Cognate Detection and Bayesian Phylogenetic Inference in Computational Historical Linguistics, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6225–6235, Association for Computational Linguistics, Florence, Italy, doi:10.18653/v1/P19-1627.

Sanda REINHEIMER RIPEANU (2001), *Lingvistica Romanica: Lexic, Morfologie, Fonetica*.

Jorma RISSANEN (1978), Modeling by Shortest Data Description, *Automatica*, 14(5):465–471, ISSN 00051098, doi:10.1016/0005-1098(78)90005-5.

Frank ROSENBLATT (1958), The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, *Psychological review*, 65(6):386–408, doi:10.1037/h0042519.

Naruya SAITOU and Masatoshi NEI (1987), The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees, *Molecular Biology and Evolution*, 4(4):406–425.

Robert. R. SOKAL and Charles. D. MICHENER (1958), A Statistical Method for Evaluating Systematic Relationships, *University of Kansas Scientific Bulletin*, 28:1409–1438.

Ilya SUTSKEVER, Oriol VINYALS, and Quoc V. LE (2014), Sequence to Sequence Learning with Neural Networks, in *Advances in Neural Information Processing Systems*, pp. 3104–3112.

Stijn Marinus VAN DONGEN (2000), *Graph Clustering by Flow Simulation*, Ph.D. thesis, University of Utrecht.

Ashish VASWANI, Noam SHAZEER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N. GOMEZ, Łukasz KAISER, and Illia POLOSUKHIN (2017), Attention Is All You Need, in I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN, and R. GARNETT, editors, *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, Curran Associates, Inc.

Andrew J. VITERBI (1967), Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm, *IEEE Transactions on Information Theory*, 13(2):260–269, ISSN 0018-9448, doi:10.1109/TIT.1967.1054010.

Hannes WETTIG, Suvi HILTUNEN, and Roman YANGARBER (2011), MDL-Based Models for Alignment of Etymological Data, in *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pp. 111–117.

Martijn WIELING, Jelena PROKIĆ, and John NERBONNE (2009), Evaluating the Pairwise String Alignment of Pronunciations, in *Proceedings of the EAACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pp. 26–34, Association for Computational Linguistics.

Shijie WU and Ryan COTTERELL (2019), Exact Hard Monotonic Attention for Character-Level Transduction, in Preslav NAKOV and Alexis PALMER, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1530–1537, Association for Computational Linguistics, Florence, Italy, doi:10.18653/v1/P19-1148.

Peter Dekker

Ⓘ 0000-0003-4734-2668
peter.dekker@ai.vub.ac.be

AI Lab
Vrije Universiteit Brussel
Pleinlaan 2
1050 Brussels
Belgium

Willem Zuidema

Ⓘ 0000-0002-2362-5447
W.H.Zuidema@uva.nl

Institute for Logic, Language
and Computation (ILLC)
University of Amsterdam
P.O. Box 94242
1090 GE Amsterdam
The Netherlands

Peter Dekker and Willem Zuidema (2020), *Word prediction in computational historical linguistics*, *Journal of Language Modelling*, 8(2):295–336

Ⓙ <https://dx.doi.org/10.15398/jlm.v8i2.268>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

ⒸⒿ <http://creativecommons.org/licenses/by/4.0/>

Serial verb constructions and covert coordinations in Edo – an analysis in Type Logical Grammar

Ralf Naumann and Thomas Gamerschlag
Heinrich-Heine-Universität Düsseldorf

ABSTRACT

Based on both syntactic and semantic criteria, Stewart (2001) and, following him, Baker and Stewart (1999), distinguish two types of serial verb constructions (SVC) and one type of covert coordination (CC) in Edo. In this article, we present an analysis of these constructions, using Type Logical Grammar (TLG) with an event-based semantic component. We choose as base logic the non-associative Lambek calculus augmented with two unary multiplicative connectives ($\text{NL}(\diamond, \square)$). SVCs and CCs are interpreted as complex event structures. The complex predicates underlying these structures are derived from simple verbs by means of a constructor. SVCs and CCs differ in terms of which part of the complex event structure is denoted. For SVCs, this is the sum of all events in the structure whereas for a CC this is only the first event in the sequence. The two verbs in an SVC and a CC are treated asymmetrically by assuming that the first verb has an extended subcategorization frame. The additional argument is of type vp (possibly modally decorated). Constraints on word order and the realization of arguments are accounted for using structural rules like permutation and contraction. The application of these rules is *enforced* by making use of the unary connectives.

*Keywords: Type
Logical Grammar,
Edo, serial verb
constructions,
covert
coordinations*

1 SERIAL VERB CONSTRUCTIONS
AND COVERT COORDINATIONS IN EDO

A standard characterization of serial verb constructions (SVCs) is (1) (Aikhenvald 2006).

- (1) An SVC is a sequence of two or more verbs with one subject and one value for tense and aspect in which the verbs are combined without overt coordination or subordination. Serial verb constructions describe what is conceptualized as a single event.

This criterion is necessary only because it is also satisfied by a similar yet distinct construction, the so-called covert coordination (CC). A common strategy to distinguish the two constructions is to use the criterion of argument sharing. For SVCs but not for CCs one has (2).

- (2) In an SVC an internal argument is shared.

SVCs occur in every language belonging to the Kwa family (Niger-Congo) like Edo, Yoruba or Igbo. They are also found in many creole languages which have a Kwa substrate, such as Haitian.

For Edo, Stewart (2001) and, following him, Baker and Stewart (1999) distinguish two types of SVCs and one type of CC.¹ In (3) each construction is illustrated by an example and the name given to the construction by Stewart (2001).² The examples below are taken from Baker and Stewart (1999:3).

- (3) a. Òzó ghá gbè èwé wù.
Ozo FUT hit goat die
'Ozo will strike the goat dead.' R SVC
- b. Òzó ghá gbè èwé khièn.
Ozo FUT hit goat sell
'Ozo will kill the goat and sell it.' C SVC

¹Baker and Stewart (2001) distinguish also a third type, a purposive SVC which will not be discussed in this article.

²In writing the Edo examples we follow Stewart (2001) and Baker and Stewart (1999) who use the standard Edo orthography (see e.g. Agheyisi 1986), adding markings of high tone (á), low tone (à) and downstep (!).

- c. Òzó ghá gbè èwé khièn ùhùnmwùn érèn.
 Ozo FUT hit goat sell head its
 ‘Ozo will kill the goat and sell its head.’ CC

This classification is based both on syntactic and semantic criteria, such as the type of the verbs, the distributional and interpretatory patterns of adverbs and the argument identifications between the verbs.

Patterns of argument identifications

1.1

In a ‘resultative serial verb construction’ (RSVC), V_1 is either transitive or intransitive whereas V_2 is either a stative, unaccusative or transitive verb with an unaccusative variant like ‘lala’ (enter).³ If V_2 is stative, V_1 is transitive. The examples below are taken from Stewart (2001).

- (4) a. Òzó kòkó Àdésúwà mósé.
 Ozo raise Adesuwa be-beautiful
 ‘Ozo raised Adesuwa to be beautiful.’ tr. + stative
 Stewart (2001:12)
- b. Òzó sùá Úyi dé.
 Ozo push Uyi fall
 ‘Ozo pushed Uyi down.’ tr. + unacc.
 Stewart (2001:8)
- c. Òzó dé wú.
 Ozo fall die
 ‘Ozo fell to death.’ unacc. + unacc.
 Stewart (2001:15)
- d. Òzó sàán kpàá.
 Ozo jump leave
 ‘Ozo jumped out.’ unerg. + unacc.
 Stewart (2001:15)
- e. Òzó gbé èkhù làá òwá.
 Ozo hit door enter house
 ‘Ozo hit the door into the house.’ tr. + tr.
 Stewart (2001:145)

³Thus, combinations of a transitive/intransitive V_1 with an unergative V_2 are excluded.

In an RSVC with a transitive V_1 and an intransitive V_2 , the only argument of V_2 is identified with the object argument of V_1 . (4a) can only mean that Adesuwa is beautiful as a result of the raising. The interpretation that Ozo became beautiful as a consequence of his raising Adesuwa is not possible. In intransitive-unaccusative pairs, both arguments are identified with each other and in the rare pattern of two transitive verbs, the direct object of V_1 is identified with the subject of V_2 .

In a ‘consequential serial verb construction’ (CSV), the verbs are either transitive or ditransitive. The subjects and direct objects are always identified with each other. By contrast, the indirect object of a ditransitive verb is never identified with any argument of the other verb. In particular, the indirect objects are not identified if both verbs are ditransitive.

- (5) a. Òzó lé èvbàré ré.
Ozo cook food eat
‘Ozo cooked food and ate it.’
Stewart (2001:60)
- b. Òzó rhié íghó hàé Úyi
Ozo take money pay Uyi
‘Ozo took some money and paid Uyi it.’
Baker and Stewart (2001:27)
- c. Úyi hàé Ìsòkèṅ íghó dó-rhié
Uyi pay Isoken money steal
‘Uyi paid Isoken the money and stole it.’
Stewart (2001:137)
- d. Òzó vbó òkhòkòhò ìgàṅ rhié nè Úyi.
Ozo pluck chicken feather give to Uyi
‘Ozo plucked the chicken of its feathers and gave them to Uyi.’
Baker and Stewart (1999:35)

The possible argument patterns for the two types of SVCs are summarized in (6).

- | | | |
|-----|-----------------------|--|
| (6) | RSVC | CSV |
| | $V_1(x) + V_2(x)$ | $V_1(x,y) + V_2(x,y)$ |
| | $V_1(x,y) + V_2(y)$ | $V_1(x,y) + V_2(x,y,z); V_1(x,y,z) + V_2(x,y)$ |
| | $V_1(x,y) + V_2(y,z)$ | $V_1(x,y,z_1) + V_2(x,y,z_2)$ |

In a CC only the subject arguments are identified whereas the object arguments do not have to be coreferential.

- (7) a. Àbié!yúwà hìín èrhán kpàán àlìmó.
 Abieyuwa climb tree pluck orange
 ‘Abieyuwa climbed the tree and plucked an orange.’
 Stewart (2001:4)
- b. Òzó gbé èkhù lá òwá.
 Ozo hit door enter house
 ‘Ozo hit the door and [he] entered the house.’
 Stewart (2001:89)

Despite the fact that the subjects are always identified, it is not possible to have a subject pronoun before V_2 in a CSV, see the example in (8a). Similarly, a subject pronoun before V_2 in an RSVC is not admissible although the subject of V_2 is identified with the object argument of V_1 (8b) (examples from Stewart 2001:64)

- (8) a. *Òzó_k mú èmà Ó_k kpèé.
 Ozo carry drum he beat
- b. *Òzó kòkó Àdésúwà_k Ó_k mósé.
 Ozo raise Adesuwa she be_beautiful

This restriction does not hold for a CC. It is possible to have a subject pronoun before V_2 , provided it is coreferential with NP_1 .

- (9) Òzó_k gbòṣ ívìn Ó_k bóló ókà.
 Ozo plant coconut he peel corn
 ‘Ozo planted coconut and [he] peeled the corn.’
 Stewart (2001:65)

If in a CC the object arguments are coreferential, there is a pronoun after V_2 that is anaphoric to NP_2 .

- (10) Òzó_k lé ízè_j Ó_k rí órè_j.
 Ozo cook rice he eat it
 ‘Ozo cooked rice and he ate it.’
 Stewart (2001:64)

Though the object arguments are always identified with each other in a CSV, it is not possible to have either an NP or a pronoun coreferential with NP_2 after V_2 . (11) cannot be interpreted as a CSV but only as a CC.

- (11) *Òzó lé ízè_k rrí ọrè_k
 Ozo cook rice eat it
 if interpreted as a CSVC, possible as a CC
 Stewart (2001:61)

From what has been said one arrives at the syntactic patterns of RSVCs and CSVCs in (12).

- (12) RSVC
 tr. + unacc./stat. NP₁ V₁ NP₂ V₂
 intr. + unacc. NP₁ V₁ V₂
 tr. + tr. NP₁ V₁ NP₂ V₂ NP₃
 CSVC
 tr. + tr. NP₁ V₁ NP₂ V₂
 tr. + ditr. NP₁ V₁ NP₂ V₂ NP₃
 ditr. + tr. NP₁ V₁ NP₂ NP₃ V₂
 ditr. + ditr. NP₁ V₁ NP₂ NP₃ V₂ NP₄

1.2 *Distribution of manner adverbs*

A last criterion that is relevant for an analysis of SVCs and CCs is the distribution of manner adverbs. Adverbs like ‘giegie’ (quickly) occur to the left of the verb and to the right of the subject and possible tense/aspect markers. They cannot occur in sentence-final position, i.e. either after the verb (intransitive verb) or the direct object (transitive verb).^{4,5}

- (13) Òzó ghá gié!gié kó!kó ọgọ (*gié!gié).
 Ozo FUT quickly gather bottle (*quickly)
 ‘Ozo will quickly gather the bottles.’
 Stewart (2001:21)

⁴Stewart (2001) as well as Baker and Stewart (1999) discuss a second type of manner adverbs the distribution of which differs from that of the adverbs discussed in the text. See Stewart (1996) for a discussion and analysis of this second class of manner adverbs.

⁵We have added the adverb in the ungrammatical position to the original example by Stewart following his observation and similar examples given by him.

A manner adverb like 'giegie' can be separated from the verb by a frequency adverb like 'ghá' (repeatedly) as in (14).

- (14) Òzó ghá gié!gié ghá kó!kó ògò.
Ozo FUT quickly ITER gather bottle
'Ozo will quickly gather the bottles repeatedly.'
Stewart (2001:21)

The schematic representation of a simple sentence is given in (15) (T/A = tense/aspect; F-Adv = frequency adverb).

- (15) simple sentence
NP₁ (T/A) (M-Adv) (F-Adv) V (NP₂) (NP₃)

For manner adverbs like 'giegie', in an RSVC the only position admissible is the one which corresponds to the position that is also admissible in a simple sentence. By contrast, CSVCs and CCs license two positions for these adverbs. Besides the position that is admissible in a simple sentence, the adverbs can also occur before the second verb. An analogous argument applies to frequency adverbs like 'ghá'. The distribution of manner adverbs like 'giegie' is shown below.

- (16) RSVC
- a. Òzó gié!gié ghá sú!á ògò dé.
Ozo quickly ITER push bottle fall
'Ozo quickly pushed the bottles down repeatedly.'
Stewart (2001:24)
- b. Òzó sùá ògò (*gié!gié) dé.
Ozo push bottle (*quickly) fall
Stewart (2001:26)
- (17) CSVC
- a. Òzó gié!gié dún!mwún èmà khién!né.
Ozo quickly pound yam sell.PL
'Ozo quickly pounded the yams and sold them.'
Stewart (2001:24)
- b. Òzó dùnmwún èmà gié!gié khién.
Ozo pound yam quickly sell
'Ozo pounded the yam and quickly sold it.'
Stewart (2001:29)

(18) CC

- a. Òzó gié!gié gbó!ó ívìn b̀l̀ó ókà.
 Ozo quickly plant coconut peel corn
 ‘Ozo quickly planted the coconut and [he] peeled the corn.’
 Stewart (2001:24)
- b. Òzó gb̀òò ívìn gié!gié bó!l̀ó ókà.
 Ozo plant coconut quickly peel corn
 ‘Ozo planted the coconut and [he] quickly peeled the corn.’
 Stewart (2001:29)

The distributional pattern of manner adverbs is summarized below.

position 1: NP₁ (T/A) Adv V₁ (NP₂) (NP₃) V₂ (NP₄)

position 2: NP₁ (T/A) V₁ (NP₂) (NP₃) Adv V₂ (NP₄)

position	1	2
RSVC	yes	no
CSVC	yes	yes
CC	yes	yes

1.3 *The semantic relation expressed by an SVC and a CC*

In an RSVC a causal relation is expressed. The first verb expresses the cause and the second verb the effect. For example, in (19), taken from Stewart (2001:13) the falling of Uyi is an effect that is triggered by the pushing, which, therefore, functions as the cause of the falling event.

- (19) Òzó sùá Úyi dé.
 Ozo push Uyi fall
 ‘Ozo pushed Uyi down.’ tr. + unacc.

In contrast to RSVCs, CSVCs and CCs do not express a causal relation. In a CSVC the relation between the two verbs is that of a consequence. The two events are ordered in the sense that the beginning point of the second event weakly succeeds the end point of the first event. In addition, e_1 is executed by the agent in order to be able to execute e_2 , i.e. e_1 is done by the agent with the eventual execution of e_2 in mind so that he can be said to follow a plan. Consider the example in (20).

- (20) Òzó lé èvbàré ré.
 Ozo cook food eat
 ‘Ozo cooked food and ate it.’
 Stewart (2001:60)

This sentence has the interpretation that Ozo cooked the rice with the intention to eat it afterwards, and, in effect, ate it. Thus, the cooking is a kind of a prerequisite for the eating so that the former is done on purpose to facilitate bringing about an event denoted by the second verb. As noted by Stewart (2001:80), the interpretation according to which Ozo had cooked the food with no intention in mind or with the intention of selling it afterwards but changed his mind later are both impossible. By contrast, no corresponding restriction on the interpretation exists for a CC. For instance, for the CC in (21), which directly corresponds to the CSVC in (20), all three interpretations are possible.

- (21) Òzó_k lé ízè_j Ó_k rrí órè_j.
 Ozo cook rice he eat it
 ‘Ozo cooked rice and he ate it.’
 Stewart (2001:64)

(21) is true in a situation in which Ozo cooked the rice with the intention to eat it and in effect ate it, in a situation where the cooking was done with no particular intention as to how to use the cooked rice but was followed by eating it, and in a situation where the cooking was done with a particular intention in mind that was not to eat it afterwards, followed by a change of mind and eating the cooked rice.

*Semantic interpretation of SVCs and CCs
 with manner adverbs*

1.4

A manner adverb in position 1 of a CC has scope only over V_1 . For example, sentence (22) means that the planting of the coconuts was quick. No corresponding assertion is made about the relative duration of the peeling of the corn. It could have been done quickly or not.

- (22) CC
 Òzó gié!gié gbó!ó ívìn b̀l̀ó ókà.
 Ozo quickly plant coconut peel corn
 ‘Ozo quickly planted the coconut and [he] peeled the corn.’
 Stewart (2001:24)

By contrast, a manner adverb in position 1 of either an RSVC or a CSVC is interpreted as modifying both verbs. (23a) is true only if both pushing and falling were quick. (23b) gets the interpretation that the whole process of pounding-plus-selling the yams was quick (compared to other pounding-plus-sellings). It says nothing about how long the pounding and selling phases take separately, compared to each other or to simple poundings and sellings (Baker and Stewart 1999:16).

- (23) SVC
- a. Òzó gié!gié ghá sú!á ògò dé.
Ozo quickly ITER push bottle fall
'Ozo quickly pushed the bottle down repeatedly.'
(Stewart 2001:24)
 - b. Òzó gié!gié dún!mwún èmà khién!né.
Ozo quickly pound yam sell.PL
'Ozo quickly pounded the yams and sold them.'
Stewart (2001:24)

If the manner adverb occurs in position 2, only V_2 is modified both for a CSVC and a CC. For (24a) to be true, the selling had to be quick whereas there is no condition on the relative duration of the pounding. Analogously, (24b) says that the peeling of the corn was done quickly but no corresponding claim is made about the planting of the coconuts.

- (24) CSVC and CC position 2
- a. Òzó dùnmwún èmà gié!gié khién.
Ozo pound yam quickly sell
'Ozo pounded the yam and quickly sold it.'
Stewart (2001:29)
 - b. Òzó gbòḡ ívìn gié!gié bó!ló ọkà.
Ozo plant coconut quickly peel corn
'Ozo planted the coconut and [he] quickly peeled the corn.'
Stewart (2001:29)

1.5

The agenda

From the discussion in this section one arrives at the following agenda of problems that have to be addressed.

- (i) How can two (or more) verbs combine with each other if that combination is realized by neither overt coordination nor overt subordination?
- (ii) How can the difference between a CSV and a CC with respect to object realization be explained? More precisely, how can we account for the fact that the object argument of a CSV cannot be overtly realized while it can be in a CC, for example by an NP or a pronoun?
- (iii) How can the distributional pattern of manner adverbs like ‘giegie’ be explained?
- (iv) How can the semantic differences between SVCs and CCs be explained?

The answers to these questions are based on the semantic interpretation of SVCs and CCs. We assume an event-based Neo-Davidsonian framework in which each verb has an additional event argument. The basic idea behind the interpretation of SVCs and CCs is that they are the result of extending an event structure made up by a single event predicate to a more complex structure with two (or possibly more) event predicates in which the events are linked by a particular relation, e.g. a causal one as in an RSVC. Such complex event structures are built by means of special constructors that operate on (the denotation of) projections of verbs. The general scheme for two transitive verbs is given in (25).

$$(25) \quad \lambda V_1. \lambda VP_2. \lambda y. \lambda x. \lambda e. \exists e_1. \exists e_2 \\ [V_1(y)(x)(e_1) \wedge VP_2(x)(e_2) \wedge \\ arg\text{-}pattern(e_1, e_2, x, y) \wedge relation(e, e_1, e_2)].$$

In (25) $arg\text{-}pattern(e_1, e_2, x, y)$ determines which arguments are shared; $relation(e, e_1, e_2)$ specifies the relation between the three events. If (25) is applied to a verb in the lexicon that can be the first verb in an SVC or a CC, one gets a complex verb which has an additional argument corresponding to the VP which specifies the sort of the event by which the event structure underlying the first verb is extended. Hence, our answer to the first question is that verbs in the lexicon can be lifted to complex predicates. Our answer to question (iv) is based on the way the events e , e_1 and e_2 are linked by $relation(e, e_1, e_2)$. In an SVC, e always is the join of e_1 and e_2 . As an

effect, manner adverbs in position 1 are interpreted relative to this complex event, yielding the interpretation that the whole action sequence has the property expressed by the adverb. By contrast, in a CC e is e_1 so that only this latter event gets modified, again in accordance with the data. Details will be given in Section 2.

Verbs in Edo that can occur as the first verb in an SVC and a CC have two different, though related subcategorization frames. The first one is the default frame assumed for canonical verbs in an SVO language. This default frame is extended by an argument of syntactic type VP if this verb occurs as the first verb in an SVC or a CC. This additional argument is looked for to the right and is the first on the subcategorization list. Proceeding in this way raises the following, further question that has to be added to the agenda.

- (v) Since the order in which the arguments of an extended verb are discharged does not coincide with the linear order in which the arguments occur in an SVC, how can the latter order be accounted for?

Questions (ii) and (v) will be answered by assuming that the logic contains a permutation and a contraction rule. This strategy is outlined in Section 3 and fully developed in Section 4. The third question will be answered by using modal decorations. This strategy makes it possible to distinguish between expressions of type A and those of type $\odot A$, where \odot is a sequence of modal operators. If modification with an adverb requires the modified expression to be of type A, the second verb in an RSVC will only project expressions of type $\odot A$ (and not of type A), whereas first verbs will have projections of the licensing type A.

The rest of the article is organized as follows. In Section 2, we introduce the semantic analysis of SVCs and CCs in Edo. Section 3 explains the basic ideas underlying the syntactic derivations of SVCs and CCs. Sections 4.1–4.3 show how the (syntactic) VP constituent in SVCs and CCs is derived. In Section 4.4, a structural rule for the subject argument is provided. In addition, the derivational semantics for CSVCs and CCs with two transitive verbs is given using examples from Section 1. In the following two sections, simple sentences with transitive verbs (Section 4.5) and simple sentences and CCs with intransitive verbs are derived (Section 4.6). Section 4.7 derives RSVCs and in Sec-

tion 4.8 we turn to the derivation of CSVs with ditransitive verbs. In Section 4.9, we sketch the analysis of manner adverbs. In Section 5, we compare our theory to those of Baker & Stewart and Ogie.

THE INTERPRETATION OF VERBS

2

Any semantic interpretation of SVCs and CCs in Edo has to take into account (i) the meaning relation between the two event predicates, and (ii) the interpretation at the level of event structure these constructions get when they are modified by a manner adverb: an adverb in an SVC can semantically have scope over both verbs in the sense that it is the joint action made up by the action expressed by V_1 and the action expressed by V_2 that is required to have the property expressed by the adverb. By contrast, in a CC a manner adverb in position 1 imposes a condition only on the action expressed by the first verb and not on the joint action.

The starting point of our analysis is the most prominent semantic characterization of SVCs: they refer to ‘single’ or ‘macro’ events. For example, as already cited in (1) and repeated in (26), Aikhenvald (2006:1) defines SVCs as follows.

- (26) A serial verb construction (SVC) is a sequence of verbs which act together as a single predicate without any overt marker of coordination, subordination or syntactic dependency of any sort. Serial verb constructions describe what is conceptualized as a single event.

Other authors using this semantic characterization include Stewart (2001), Baker and Stewart (1999) and Dixon (2006). One problem with this definition is that the notion of a single or a macro event needs to be made precise. Consider first the example in (27) from Yimas, a Papuan language of new Guinea, taken from Foley (2010:81).⁶

⁶ OBL: oblique; VIII: noun class 8; SG: singular; O: other argument; A: agent-like participant; SEQ: sequential.

- (27) a. arm-n kay
 water-OBL canoe-VIII-SG
 i-ka-ak-mpi-wul.
 VIII-SG-O-1SG-A-push-SEQ-put-in
 ‘I pushed the canoe down into the water.’

This sentence is an SVC since it is monoclausal and the pronominal agreement affixes must precede the sequence of verbs.⁷ However, Foley argues that ‘ak-mpi-wul’ (push down into [the water]) does not denote a single event. It rather refers to ‘one (or more commonly, multiple) actor(s) causing a canoe to move linearly along the ground away from the high ground of the riverbank toward the lower level of the river itself, so that it descends down the edge of the riverbank and comes to float on the water of the river’, Foley (2010). One may counter this argument by requiring that by a ‘single’ or a ‘macro’ event is not necessarily meant an atomic event but possibly a complex event that can have other events as material or mereological parts. This move, however, immediately raises the following problem discussed in Bohnemeyer *et al.* (2007). If one assumes that the domain of events is structured by a material part-of relation \sqsubseteq and a sum operation \sqcup in the sense of Link (1998), and given that the interpretation of an expression requires the existence of n events e_1, \dots, e_n , then there always exists the sum event $e = e_1 \sqcup \dots \sqcup e_n$. Bohnemeyer *et al.* (2007:500) illustrate this problem with the following minimal pair taken from English and Ewe, a Gbe language of the Kwa family within Niger-Congo that is spoken in Ghana and Togo.⁸

- (28) The circle rolled from the blue square past the house-shaped object to the green triangle.
- (29) Circle lá mli tsó blutɔ gbó le mɔ́-á dzí tó
 circle DEF roll from blue place LOC road-DEF top pass
 xɔ-a ɲú yi dɛ́ triangle lá gbó.
 house-DEF skin go ALL triangle DEF place
 ‘The circle rolls from the blue place on the road, passes the side of the house, goes to the triangle.’

⁷ Foley (1991) argues that it is in effect a single grammatical word.

⁸ In the examples below one has: DEF: definite; LOC: locative; ALL: allative.

Whereas in English a single VP is sufficient, Ewe requires three. Does this mean that in English only a single, though complex, event is described whereas in Ewe three events are described? Given a domain of events structured by a part-of and a sum operation, there always is a sum of three events in addition to the three events of rolling, passing and going-to so it is always possible to claim that the whole clause in (29) is interpreted relative to this sum. As a result, both options are at least theoretically possible. One attempt at solving this problem is to assume that if a clause contains n event predicates, each predicate is interpreted relative to the sum of the n events. For (29), this amounts to interpreting each of the three event predicates relative to the sum event consisting of a rolling, a passing and a going-to event. However, this strategy fails for the following reason. An atomic event predicate P is always interpreted relative to (sums of) events of the same sort, e.g. a rolling or a passing but not relative to 'heterogeneous' events, for example sums of rollings and/or passings. From this it follows that each event predicate in a clause has to be interpreted relative to a (sum) event that is the join of events of the same sort. For example, in the Ewe example above 'mli' (roll) has to be interpreted relative to (sums of) rolling events, 'tó' (pass) has to be interpreted relative to (sums of) passing events, and 'yi' (go) has to be interpreted relative to (sums of) going (to) events. Hence, in order to be true, any clause containing n event predicates requires the existence of n 'homogeneous' events in relation to which the n predicates are interpreted. Using a structured domain of events, this existence implies the existence of a corresponding sum event which consists of n homogeneous events. Since these n events belong to different sorts, this sum is heterogeneous.

The above discussion tried to locate the difference between SVCs and other multi-verb constructions at the ontological level, i.e. at the level of real-world events. In contrast to this failed strategy, Bohne-meyer et al. propose to locate this distinction at the level of constructions. Specifically, they take this difference to be located at the level of the form-to-meaning property of event descriptions. They define this property, the macro event property (MEP), by reference to temporal operators:

DEFINITION 1 Let expression C denote an event predicate P ($\llbracket C \rrbracket = \exists e.P(e)$). Let T_{POS} be any modifier of C ($\llbracket \dots T_{POS} \dots \rrbracket_C$) that locates some subevent $e' \sqsubseteq e$ at time t ($\llbracket T_{POS} \rrbracket = \lambda Q.\lambda t.\exists e'[Q(e') \wedge \tau(e') \subseteq t]$, where Q may or may not be identical to P). Then C has the macro-event property (MEP) iff any syntactically and semantically acceptable T_{POS} necessarily also locates e at t (i.e. $AT(Q, e', t) \rightarrow AT(P, e, t)$ for any acceptable T_{POS} and $AT := \lambda P.\lambda t.\exists e(P(e) \wedge \tau(e) \subseteq t)$).

Intuitively, an expression or construction has the MEP if it licenses only temporal operators that have scope over all subevents, (Bohnmeyer *et al.* 2007:507). Note that the MEP does not make any assertion about the kinds of events a construction having the MEP can refer to. In particular, no ontological type of ‘macro-event’ is singled out or presupposed that can be distinguished from other, non-macro events. The English example in (28) trivially has the MEP because there is only one event predicate in the VP. For the Ewe example in (29) the MEP follows from the fact that any time-positional operator must have scope over all three VPs. Modifying all three VPs separately with a time adverbial leads to ungrammaticality, see (30) taken from Bohnmeyer *et al.* (2007:506).

- (30) *Circle lá mli tsó blutɔ gbó le mó-a dzí le ga
 circle DEF roll from blue place LOC road-DEF top at hour
 enyí me tó xɔ-a ɲú le ga asiéke me yi dé
 eight in pass house-DEF skin at hour nine in go ALL
 triangle lá gbó le ga ewó me.
 triangle DEF place at hour ten in
 Intended: ‘The circle rolls from the blue place on the road at
 eight o’clock, passes the side of the house at nine ’clock, goes
 to the triangle at ten o’clock.’

Bohnmeyer *et al.* (2007) discuss an additional example from English (The sentences in (31)–(34) are taken from Bohnmeyer *et al.* 2007).

- (31) Floyd went from Rochester via Batavia to Buffalo in the morning.

In (31) ‘in the morning’ modifies the whole motion event including the departure, the passing and the arriving. The time adverbial used must be of the appropriate sort. Since (31) refers to an event with an extended run-time, adverbials denoting a time point are excluded.

- (32) ?Floyd went from Rochester via Batavia to Buffalo at seven/eight-thirty.

Trying to ‘time’ the corresponding phases leads to ungrammaticality, see (33).

- (33) *Floyd went from Rochester at seven via Batavia at seven forty-five to Buffalo at eight thirty.

If one wants to modify the three phases separately, one has to use different verbs for the departure, the passing and the arrival as in (34).

- (34) Floyd left Rochester at seven, passed through Batavia at seven forty-five, and arrived at Buffalo at eight thirty.

As it stands, the MEP only applies to temporal modifiers. Foley (2010) generalizes the MEP to other kinds of modifiers. According to him, the MEP requires that temporal operators, adjuncts, adverbial clauses and tense affixes have scope over all component sub-events that are denoted by event predicates in the construction. How can this modification be incorporated into an event-based framework? Foley’s generalization shows that the MEP can be applied to various properties of events like their run-time or the speed with which they are executed. In a standard event semantics such properties are uniformly interpreted as sets of events, similarly to sortal distinctions like poundings and sellings. We have to leave open the question to which dimensions in a particular language the MEP can apply. For Edo, one dimension is that of speed for which the adverb ‘giegie’ specifies a particular value. A second important question that has to be left open is: is it possible that two modifiers differ with regard to the MEP in the sense that one imposes the MEP whereas the other does not?

The MEP in Edo

2.1

In this section we will adapt the results of the discussion in the previous section to Edo. In Bohnemeyer et al.’s account the mapping is guided by the interpretation of temporal operators. If such an operator has scope over all event predicates, the whole construction has the MEP. Applied to Edo, a weakness of this analysis is that it is not related to the semantic interpretation of the whole construction in the

sense that no reference is made to the meaning relation that holds between the event predicates in the construction. In contrast to this way of defining the MEP, we will base our analysis on the semantic relation expressed by SVCs and CCs. Recall that both in an RSVC and a CSVC the two events are not only related at the temporal level by a weakly succession relation but there is an additional non-temporal relation that holds between the two events: a causal relation in the case of an RSVC and a plan (intention) relation in the case of a CSVC. One way of looking at an SVC from this perspective is to analyze it as something built from a complex predicate constructor that maps two (or possibly more) event predicates to a complex predicate. This process is constrained both at the level of shared arguments (argument pattern) and at the level of how the events are related to each other.⁹ A scheme of such a constructor for two event predicates is given in (35).

$$(35) \quad \lambda P_1. \lambda P_2. \lambda y. \lambda x. \lambda e. \exists e_1. \exists e_2 [P_1(e_1) \wedge P_2(e_2) \\ \wedge \text{arg-pattern}(e_1, e_2, x, y) \wedge \text{relation}(e, e_1, e_2)].$$

P_1 and P_2 are two event predicates that correspond to V_1 and V_2 in a complex predicate, respectively. *arg-pattern* and *relation* are parameters whose value depends on the type of the complex predicate (CSVC, RSVC or CC). *arg-pattern*(e_1, e_2, x, y) is the constraint on the argument pattern while *relation*(e, e_1, e_2) is the constraint on the relation between the events. For example, for the CSVC in (20), *arg-pattern* identifies both the actors and the themes of the events related to P_1 and P_2 .¹⁰ The result is a complex predicate whose subcategorization frame is that of the (identical) subcategorization frames related to the two event predicates. For the relation between the events, in particular the definition of \square_x , see below for details.

$$(36) \quad \lambda y. \lambda x. \lambda e. \exists e_1. \exists e_2 [\text{cook}(e_1) \wedge \text{eat}(e_2) \wedge \text{actor}(e_1) = x = \\ \text{actor}(e_2) \wedge \text{theme}(e_1) = y = \text{theme}(e_2) \wedge e = e_1 \sqcup e_2 \wedge e_1 \preceq \\ e_2 \wedge \square_x(\text{occur}(e_1) \rightarrow \text{occur}(e_2))].$$

⁹The use of the word ‘constructor’ must not be misunderstood as referring to some form of construction grammar. Rather, it refers to the fact discussed and explained below that it is an operation which builds a complex event structure out of a simple one.

¹⁰*arg-pattern* and *relation* will be discussed below.

In (37) the case of the RSVC in (19) is given. In this case the argument pattern identifies the theme arguments of e_1 and e_2 whereas the actor of e_1 remains unrelated. *relation* requires the two events to be causally related (see below for details).

$$(37) \quad \lambda y. \lambda x. \lambda e. \exists e_1. \exists e_2 [push(e_1) \wedge fall(e_2) \wedge actor(e_1) = x \wedge theme(e_1) = y = theme(e_2) \wedge e = e_1 \sqcup e_2 \wedge cause(e_1, e_2)].$$

The constructor in (35) applies only to cases where all arguments related to the second event predicate are shared with an argument related to the first event predicate. At first sight this might be problematic for SVCs in which not all arguments are shared because then non-shared arguments would have to be added as arguments to the resulting complex predicate, which empirically is not the case. Recall that non-shared arguments (related to the second event predicate) are allowed in a CSVC with two ditransitive verbs where the indirect objects must be different, in an RSVC with two transitive verbs and in a CC where no constraints are imposed on the direct objects. This lack of generality stems from the fact that both event predicates are taken on a par. Rather, one has to view the complex predicate constructor as a way to extend an event structure comprising only one event predicate to a more complex event structure that contains two (or possibly more) event predicates and in which the events are related by particular constraints. What gets extended is always the event predicate whose corresponding event is executed first in the resulting event structure. The second event structure is not arbitrary. For example, both in an SVC and a CC the actors are required to be the same. A similar generalization across constructions is not possible for direct and indirect arguments. These conditions have to be reflected at the syntactic level. Instead of P_2 , the projection VP_2 of the corresponding verb V_2 has to be taken as an argument. Hence, V_2 is already partially saturated when it enters the constructor. Similarly, to make sure that the argument structure of the complex predicate is that of the first verb, we have to use V_1 instead of P_1 . Argument sharing is then expressed in terms of constraints on the respective arguments. The result for two transitive verbs is the constructor (scheme) in (38).

$$(38) \quad \lambda V_1. \lambda VP_2. \lambda x. \lambda y. \lambda e. \exists e_1. \exists e_2 [V_1(y)(x)(e_1) \wedge VP_2(x)(e_2) \wedge arg_pattern(e_1, e_2, x, y) \wedge relation(e, e_1, e_2)].$$

Let us next turn to the relation between e , e_1 and e_2 . Our central thesis is that in Edo this relation depends on the (semantic) relation that holds between e_1 and e_2 .

- (39) If the relation between e_1 and e_2 cannot be reduced to a purely temporal one, one has $e = e_1 \sqcup e_2$, otherwise one gets $e = e_1$.

The rationale behind (39) is the following. The unextended verb corresponding to an extended one expresses only one action (e_1) without taking into consideration what actions (events) can follow this first action. Extended verbs are one way of extending verbs expressing a single action to more complex sequences of actions. Hence, the cognitive significance of extending a single event predicate to a complex one is just to express this relation between the two events. This relation should therefore be reflected in the complex predicate by letting the abstracted event variable refer to the sum of the two events. By contrast, in a CC the two events are related only at the temporal level (but see below for a revised view). In this case the event input to the complex predicate is the first event similar to the case of the unextended verb form. The sum event is not needed for this temporal succession. Compare this with the sequencing operation $\alpha; \beta$: do first α and then β where the two actions need only be related at the temporal level. Hence, in an SVC, e has to be $e_1 \sqcup e_2$. By contrast, in a CC e is e_1 because it is the first event in the sequence and there is no additional relation linking the two events except the temporal one.

Furthermore, the temporal relation between e_1 and e_2 in all three kinds of complex predicates is that of weakly succeeding, denoted by \preceq : $e \preceq e'$, which holds if the beginning point of e' follows shortly after the end point of e . This condition requires that no other events involving the direct object occur between e_1 and e_2 which makes the occurrence of e_2 unlikely. For example, if Ozu killed the goat in order to sell it, he must not have eaten its meat afterwards because this makes selling it impossible. For a CC, the temporal relation is the only condition on the two events. For a CSVC, a second condition requires that the two events are part of a common plan. This condition is modelled by $\Box_x(\text{occur}(e_1) \rightarrow \text{occur}(e_2))$, which requires that in all worlds that are compatible with what the agent x plans to do an occurrence of e_1 implies an occurrence of e_2 . For an RSVC, the two events are related by the relation *cause*.

So far we assumed that the thematic roles of shared arguments match. Since this assumption may turn out to be too strong, we will formulate the condition on the argument pattern in terms of a thematic role hierarchy relative to the subcategorization frame of the two verbs. Since extended verbs extend the first verb, the thematic roles of this verb are known so that the actual roles can be used. One possible thematic role hierarchy is given by Actor > Goal/Source > Theme (Grimshaw 1990). $TR(e_1) = n\text{-th}(e_2)$ is true if the object assigned by the thematic role TR to e_1 is identical to the object that is assigned to e_2 by the n -th thematic role in the thematic role hierarchy restricted to those roles that are defined in its subcategorization frame. Specifically, we assume the following patterns for two transitive verbs (CSVC and CC) and an RSVC with a transitive first and an intransitive second verb.

- CSVC : $actor(e_1) = first(e_2) \wedge theme(e_1) = second(e_2)$
- RSVC : $theme(e_1) = first(e_2)$
- CC : $actor(e_1) = first(e_2)$

We are now finally ready to give the meanings of verbs in an SVC and a CC. There are two strategies as to how the meaning of verbs that occur as first verbs in an SVC or a CC can be derived on the basis of a constructor. In the first strategy one explicitly derives the meaning from a constructor by applying this constructor to the meaning of a verb in a simple sentence. Such an operation can be performed either in the lexicon or at some later stage, say, during the derivation of an SVC or a CC. In the second strategy these meanings are not derived by an operation but rather, the result of applying one of the constructors to the meaning of a verb is taken as an additional meaning of the verb. We choose the second strategy because it is in accordance with the lexicalist assumption underlying TLG. (See below for details on how the lexicon in Edo is structured in our approach). In (40), the meaning of a CSVC with two transitive verbs and in (41), the meaning of an RSVC with a transitive and an unaccusative verb are given. In both cases, P_1 is the actual verb, for example ‘cook’ in a CSVC or ‘hit’ in an RSVC. Since these verbs have an additional argument of type VP , they will be called ‘extended verb (forms)’.

$$(40) \quad \lambda VP_2. \lambda y. \lambda x. \lambda e. \exists e_1. \exists e_2 [e = e_1 \sqcup e_2 \wedge P_1(e_1) \wedge VP_2(x)(e_2) \wedge actor(e_1) = x = first(e_2) \wedge theme(e_1) = y = second(e_2) \wedge e_1 \preceq e_2 \wedge \Box_x(occur(e_1) \rightarrow occur(e_2))].$$

$$(41) \quad \lambda VP_2. \lambda y. \lambda x. \lambda e. \exists e_1. \exists e_2 [e = e_1 \sqcup e_2 \wedge P_1(e_1) \wedge VP_2(y)(e_2) \wedge theme(e_1) = y = first(e_2) \wedge e_1 \preceq e_2 \wedge cause(e_1, e_2)].$$

(42) presents the extended verb form for a CC with two transitive verbs. Similarly to the examples of CSVs and RSVs, P_1 is the actual verb, e.g. ‘cook’.

$$(42) \quad \lambda VP_2. \lambda x. \lambda y. \lambda e. \exists e_1. \exists e_2 [e = e_1 \wedge P_1(e_1) \wedge VP_2(x)(e_2) \wedge actor(e_1) = x = actor(e_2) \wedge e_1 \preceq e_2].$$

The meanings of (first) verbs in complex predicates can be taken as a formal rendering of Foley’s insight. SVCs are interpreted relative to macro events whose component events are used in the interpretation of the atomic event predicates out of which the complex predicate is built. Interpreting SVCs relative to complex (macro) events has been suggested before (see Bohnemeyer *et al.* 2007 for an overview). However, these proposals are mostly not formalized. In particular, the exact relation between the complex event and the events denoted by the component event predicates remains unspecified.

We base our analysis on the fact that an SVC denotes a complex event structure that is built from an atomic event structure in order to express a complex action based on plans or causal relations. In what sense does this interpretation apply to CCs? Or, to put it differently: what is the cognitive or semantic significance of a CC compared to a construction that is made up by two separate sentences? In order to answer this question one has to look at the discourse level. At this level a sequence of sentences need not only be free of semantic anomalies (and be true) but in addition it has to be coherent. This means that two sentences have to be related by a coherence relation like narration, background or result. Viewed from this perspective, the thesis is that a CC and likewise an SVC are devices to build-in a coherence relation between two (or more) event predicates. For a CC, the coherence relation is that of narration. The two events are related by the temporal relation of weak succession and the two events must have a common actor. Hence, the condition for narration is satisfied (Asher and Lascarides 2001). The relation to the notion ‘Question under Discussion’

is the following. Given a context c with an event e and objects o_1, \dots, o_n participating in e , a set of implicit questions related to the event and the objects is raised. In order for a continuation of this context to cohere with this context at least one of these questions needs to be answered in the continuation. Examples of questions are ‘What next?’ at the event level and ‘What about x ?’ at the level of objects. In SVCs in Edo, these questions are further restricted. First, the events must be related by a plan or a causal relation, and, second, the next sentence must involve the same actor and the same theme, i.e. it provides further information on both objects. Hence, SVCs do by the way they are constructed answer QuDs so that, in effect, the text is coherent.¹¹

How do the meanings of verbs that occur as the first verb in a complex predicate relate to the lexicon? In TLG each lexical item is assigned a set of syntactic types. If this set is a singleton, the grammar is called *rigid*. If a lexical item is assigned more than one type, this reflects the fact that it can occur in different syntactic contexts with different types of arguments. An example in English is ‘know’ which can have an argument of type np (‘know the answer’) or a clause-like argument (‘know that p ’). Similarly, a verb in Edo is in general assigned more than one syntactic type. Which types are assigned to a verb depends on the way it can be used in SVCs and CCs. Since there are three constructions (RSVC, CSVC and CC), one gets a maximal number of four different types. The maximal number is obtained if a verb can occur as V_1 in all three constructions (three types) plus the type it is assigned in simple sentences and as V_2 in any of the three constructions. In practice, the number is smaller. For example, in an RSVC, V_1 cannot be ditransitive and in a CSVC intransitive verbs are excluded as V_1 .

Let $\sigma(\textit{verb})$ be the set of syntactic types assigned to the verb *verb*. Each element of $\sigma(\textit{verb})$ is paired with a typed λ -term as the meaning of *verb*. In Edo, one λ -term corresponds to the case of a verb in a simple sentence or as V_i with $i \geq 2$ in a complex predicate, if admissible. Other possible λ -terms result if one of the constructors is applied to the ‘standard’ λ -term as argument. Importantly, this application is *not* part of the lexicon, as already said above. Rather, only the result-

¹¹ See Naumann and Petersen (2019) for a formal theory of QuDs in a dynamic semantics with frames.

ing λ -term is. Examples are (40), (41) and (42). We leave open the question whether it is desirable to view the lexicon in Edo in such a way that at first only the meanings of simple verb forms are given and the complex meanings are derived, if admissible, by applying lifts to the meanings of these simple forms.

Let us summarize the results of this section. Both SVCs and CCs are analyzed in terms of extended verbs that are taken to be the result of applying a complex predicate constructor to an (unextended) verb. This interpretation is driven by the fact that the semantic (or cognitive) function of these constructors is to express complex event structures. The events denoted by such structures are related by particular constraints like (i) ‘What actions are successively executed by an actor?’, (ii) plans that are made up by a series of consecutive actions, and (iii) causal relations. Common to both types of construction is a built-in coherence relation (narration).

3 A GRAMMATICAL ARCHITECTURE FOR EDO IN TYPE LOGIC GRAMMAR

In this section, we will introduce the logical architecture to be used in our analysis of the Edo data presented in Section 1. The theoretical framework is a multimodal variant of the non-associative Lambek calculus NL enriched with two unary connectives.¹² For many linguistic applications, the operations available in NL are too restrictive to account for the variety of phenomena found in natural languages. For example, the only way to combine two linguistic resources consists in concatenating them, and in addition NL imposes a rigid binary constituent (or dependency) tree structure. Extending NL with the structural rules in (43) leads to overgeneration.

- (43) a. $A \bullet B \rightarrow B \bullet A$ permutation [P]
 b. $A \rightarrow A \bullet A$ contraction [C]
 c. $(A \bullet B) \bullet C \rightarrow A \bullet (B \bullet C)$ associativity [Ass]

¹²See Moot and Retoré (2012) for a more detailed introduction to multimodal calculi with unary connectives on which our presentation is based, as well as Morrill (2011).

For instance, if permutation and associativity are globally available, not only the grammatical ‘John dedicated the book to Bill’ but also the ill-formed ‘John dedicated to Bill the book’ becomes derivable. Simply substitute in the derivation below ‘the book’ for x and skip the application of the $[/I]$ rule.

$$\begin{array}{c}
 \frac{\text{dedicate} \Rightarrow \text{vp/pp/np} \quad [x \Rightarrow \text{np}]^1}{(\text{dedicate } x) \Rightarrow \text{vp/pp}} \quad [/\text{E}] \quad \text{to Bill} \Rightarrow \text{pp} \quad [/\text{E}] \\
 \frac{\text{John} \Rightarrow \text{np} \quad ((\text{dedicate } x) \text{ to Bill}) \Rightarrow \text{vp}}{(\text{John } ((\text{dedicate } x) \text{ to Bill})) \Rightarrow \text{s}} \quad [/\text{E}] \\
 \frac{(\text{John } ((\text{dedicate } x) \text{ to Bill})) \Rightarrow \text{s}}{(\text{John } ((\text{dedicate to Bill}) x)) \Rightarrow \text{s}} \quad [\text{P}] \\
 \frac{(\text{John } ((\text{dedicate to Bill}) x)) \Rightarrow \text{s}}{((\text{John } (\text{dedicate to Bill})) x) \Rightarrow \text{s}} \quad [/\text{Ass}] \\
 \frac{((\text{John } (\text{dedicate to Bill})) x) \Rightarrow \text{s}}{(\text{John } (\text{dedicate to Bill})) \Rightarrow \text{s/np}} \quad [/\text{I}]^1
 \end{array}$$

What is required is a controlled access to the device of structural rules in the sense that their application is restricted to the appropriate (licensing) contexts. One way to achieve this consists in using a *multi-modal* variant of the base logic NL. Instead of a single family $\{/, \bullet, \backslash\}$ of connectives, one distinguishes different such families: $\{/, \bullet_i, \backslash_i\}$, $i \in I$. The elements of the index set I are called *modes of combination* or simply *modes*. Each family comes with its own set of structural rules. The main function of such modes is to license or inhibit the use of structural rules only in particular contexts and to exclude it in all other contexts. Formally, the use of modes can be seen as the use of a combined logic, which is built of several subsystems, one for each mode. Underlying this strategy is the intuition that linguistic resources belonging to distinct types can have different properties. Distinguishing various modes of combination makes it possible to discern linguistic contexts that differ with respect to their properties. In each context, the same logical rules governing the operators hold. However, they possibly differ with respect to the structural rules that can be applied to them.

The various modes can be related by inclusion and interaction rules. Inclusion rules relate different modes with each other. For example, if mode $/_i$ includes mode $/_j$ and one has $A/_iB$, $A/_jB$ can be derived. An example given by Moot and Retoré (2012) is the following. If a formula of type $A/_iB$ can select its B argument both to the right and to the left as in LP, e.g., one also has $A/_jB$ relative to L, in which arguments can only be chosen to the right. Adding structural rules via

a particular mode *enables* the application of this rule but it does not *enforce* it. Therefore an observation can be made that although the formulation of structural rules in the context of a multimodal system makes it possible to restrict their application to the intended contexts, it does not force their application in these contexts.

This problem can be solved by extending the base logic in a further direction. This extension consists in adding unary operators \diamond and \square . Similarly to the family $(\bullet, /, \backslash)$, the two operators are related by a law of residuation, which is given in (44a). From this law the relationships in (44b) are derivable.

- (44) a. $\diamond A \vdash B$ iff $A \vdash \square B$
 b. $\diamond \square A \vdash A$ and $A \vdash \square \diamond A$

Analogous to the binary operators, it is possible to have a multimodal system for these unary operators. Given an index set J , one distinguishes various families of residuated pairs $\{\diamond_j, \square_j\}$ with $j \in J$. Modal decorations are primarily used in the type assignments of lexical items and in interaction rules with binary connectives, i.e. so-called K-rules. When taken together, these two strategies can be used to solve the problem of enforcing the application of a structural rule. Let us illustrate this with an example.

- (45) a. $K: \diamond_j (A \bullet_i B) \rightarrow \diamond_j A \bullet_i \diamond_j B$
 b. $K2: \diamond_j (A \bullet_i B) \rightarrow A \bullet_i \diamond_j B$

The rule K distributes \diamond_j over both components of \bullet_i , whereas K2 does this only for the right component. The relationship between the problem of enforcing the application of a structural rule in an intended context and the percolation (or distribution) of structural (modal) operators is the following. The percolation mechanism that passes a modal decoration from some substructure to a structure that is of an undecorated designated type has to be construed in such a way that it requires the application of the structural rules. Thus, structural rules are used to create contexts which license the percolation of modal decorations which are not possible if these rules are not applied.

Next we will sketch how the above architecture will be used in our analysis of the data in Edo described in Section 1. Recall that we assume that in SVCs and CCs a verb form is used that extends the

subcategorization list by an additional argument of type *vp*. One way in which CSVCs differ from CCs with two transitive verbs is that in the former construction the direct objects are always identified with each other and that the direct object of V_2 must not be overtly realized, say, by a pronoun. Hence, SVCs and CCs differ in the way direct objects are treated and in a CC the subject is treated differently from the direct object: whereas the former are always identified, this need not be the case for the direct objects. This suggests to distinguish, first, between the way subjects combine with a VP and the way a (transitive or ditransitive) verb combines with its direct object, and, second, between two head adjunction modes for the combination of an extended verb with the additional VP argument in an SVC and a CC, respectively. This yields the modes in (46) for Edo.

- (46) a. \cdot_{1l} : head-(left) complement mode (verb object relation)
 b. \cdot_{1r} : head-(right) complement mode (verb subject relation)
 c. \cdot_i : head adjunction mode for $i = 0$ or $i = 2$ (verb additional argument relation in an SVC and a CC)

Let us next illustrate an interaction rule which is a restricted form of permutation. If the extended (transitive) verb combines with the (additional) argument of type *vp* in an SVC or a CC, and then with the direct object, the order of the two arguments has to be changed. This is achieved by the mixed permutation rule in (47).

- (47) MP: $(A \bullet_{1l} B) \bullet_i C \rightarrow (A \bullet_i C) \bullet_{1l} B$
 (The subscript \cdot_i is a head adjunction mode)

This rule requires a context in which two verbal elements forming a cluster $(A \bullet_i C)$ are composed with a nominal element (B), which is to the right of the cluster. The requirement on the left component to be a verbal cluster makes this rule applicable only in the context of an SVC and a CC. Hence, this structural rule has only a controlled access to lexical resources. Now consider the following example in which the complex VP = $V_1 NP_2 V_2 NP_3$ of a CC is derived (\cdot_2 the head adjunction mode for a CC), using both logical rules (elimination rules for two constructors $/_{1l}$ and $/_2$) and the mixed permutation rule in (47).

$$\begin{array}{c}
 \text{G} \\
 \frac{v_1 \Rightarrow \text{tv}/_2 \text{vp} \quad \frac{v_2 \Rightarrow \text{tv} \quad \text{np}_3 \Rightarrow \text{np}}{v_2 \circ_{1l} \text{np}_3 \Rightarrow \text{vp}} \text{ [}_{1l}\text{E]} \\
 \frac{\quad}{(v_1 \circ_2 (v_2 \circ_{1l} \text{np}_3)) \Rightarrow \text{tv}} \text{ [}_{2}\text{E]} \quad \text{np}_2 \Rightarrow \text{np} \text{ [}_{1l}\text{E]} \\
 \frac{\quad}{\frac{((v_1 \circ_2 (v_2 \circ_{1l} \text{np}_3)) \circ_{1l} \text{np}_2) \Rightarrow \text{vp}}{((v_1 \circ_{1l} \text{np}_2) \circ_2 (v_2 \circ_{1l} \text{np}_3)) \Rightarrow \text{vp}} \text{ [MP]}
 \end{array}$$

Applying MP in line 4 yields the correct word order: NP₂ is adjacent to V₁ and precedes the additional argument, which is VP₂ = V₂ NP₃. But if MP is not applied, one rests with the sequent in line 4, which does not have a grammatical word order. This problem will be solved by introducing unary connectives. As already said above, modal decorations are used both in the type assignment to lexical items as well as in relation to the designated types, which are vp and s in our analysis. There are two different ways of how lexical items are modally decorated in our analysis: $\diamond_j \square_j A$ or $\square_j A$. The former is used as the lexical type assignment to verbs and the latter for lexical resources of type np. We start with the case of verbs. The type assignment $A: \diamond_j \square_j A$ holds both for the unextended and the extended form. Hence, verb forms used in SVCs and CCs do not differ at the level of modal decoration but at the level of the mode of combination. On the assignment $\diamond_j \square_j A$, one starts a derivation with an identity axiom $\square_j A \Rightarrow \square_j A$. Application of the logical rule $[\square_j \text{E}]$ yields the sequent $\langle \square_j A \rangle^j \Rightarrow A$. Hence, this lexical resource can function as a being of type A. $\langle \square_j A \rangle^j$ eventually becomes part of a larger constituent Γ . In our application, Γ is either VP_1 or the complex predicate consisting of VP_1 and VP_2 . Finally, $\langle \square_j A \rangle^j$ gets substituted by the lexical resource of type $\diamond_j \square_j A$ using the identity axiom $\diamond_j \square_j A \Rightarrow \diamond_j \square_j A$ and an application of the logical rule $[\square_j \text{E}]$. The derivation is schematically presented below.

$$\begin{array}{c}
 \frac{\square_j A \Rightarrow \square_j A}{\langle \square_j A \rangle^j \Rightarrow A} \text{ [}_{\square_j}\text{E]} \\
 \vdots \\
 \Gamma [\langle \square_j A \rangle^j] \Rightarrow C \\
 \vdots \\
 \frac{\Gamma [\langle \square_j A \rangle^j] \Rightarrow C \quad \diamond_j \square_j A \Rightarrow \diamond_j \square_j A}{\Gamma [\langle \diamond_j \square_j A \rangle] \Rightarrow C} \text{ [}_{\diamond_j}\text{E]}
 \end{array}$$

Hence, for extended verbs, which are modally decorated by $\diamond_j \square_j$, the modal decoration must not be removed. As will be shown next, this is different for lexical resources of type np. In the lexicon, they get the type assignment $\square_j A$. Similarly to the case of verbs, a derivation starts with an identity axiom $\square_j A \Rightarrow \square_j A$, followed by the application of the logical rule $[\square_j E]$ yielding the sequent $\langle \square_j A \rangle^j \Rightarrow A$. Again similarly to the case of verbs, the modally decorated type eventually becomes part of a larger constituent Γ , which is either VP_1 or the complex predicate consisting of this VP and the additional argument of type vp. In contrast to the use of the modal decoration for verbs, the modal decoration for NPs must be removed. Otherwise, no lexical substitution would be possible because there are no lexical resources of type $\diamond_j \square_j A$ for $A = np$. This removal is achieved by K-rules. If the NP corresponds to the direct object of V_1 , two K-rules have to be applied. The first percolates the modal decoration to VP_1 and the second to the complex predicate, say $\Gamma: \langle \Gamma[A] \rangle^j \Rightarrow C$. To this sequent, rule $[\square_j I]$ is applied, yielding $\Gamma[A] \Rightarrow \square_j C$. If $C = vp$ or $C = s$, the task consists in deriving expressions of type $\square_j vp$ and $\square_j s$ and not the corresponding non-decorated types. This is the second principle use of modal decorations. A schematic derivation is represented below.

$$\frac{\square_j A \Rightarrow \square_j A}{\langle \square_j A \rangle^j \Rightarrow A} [\square_j E]$$

$$\vdots$$

$$\Gamma [\langle \square_j A \rangle^j] \Rightarrow C$$

$$\vdots$$

$$\frac{\langle \Gamma [\square_j A] \rangle^j \Rightarrow C}{\Gamma [\square_j A] \Rightarrow \square_j C} [\square_j I]$$

The use of both kinds of modal decorations is illustrated by the following example. Consider the sequent in (48), which is a result of applying an extended verb in a CSVC or a CC to the additional vp-argument and its direct object (in that order).

$$(48) \quad (\langle \square_j A \rangle^j \circ_k \Gamma) \circ_i \langle \square_j B \rangle^j \Rightarrow C$$

Next, a rule of permutation needs to be applied in order to arrive at the correct word order. This can be achieved by the permutation rule in (49), which generalizes the corresponding rule in (47).

$$(49) \quad \text{MP: } (C \bullet_i D) \bullet_k E \rightarrow (C \bullet_k E) \bullet_i D$$

Next, the modal decoration of the B-resource must be percolated to VP_1 ($= (\langle \Box_j A \rangle^j \circ_i \langle \Box_j B \rangle^j)$). This is achieved by the K-rule in (50).

$$(50) \quad \Diamond_j (\Diamond_j A \bullet_i B) \rightarrow \Diamond_j A \bullet_i \Diamond_j B.$$

Note that this rule does not remove the modal decoration of the (verbal) resource A. Otherwise, no lexical substitution would be possible after removing the decoration. The derivation looks as follows.

$$\frac{\frac{\langle \Box_j A \rangle^j \circ_k \Gamma \circ_i \langle \Box_j B \rangle^j \Rightarrow C}{\langle \Box_j A \rangle^j \circ_i \langle \Box_j B \rangle^j \circ_k \Gamma \Rightarrow C} \text{ [MP]}}{\langle (\Box_j A \rangle^j \circ_i \Box_j B) \rangle^j \circ_k \Gamma \Rightarrow C} \text{ [(50)]}$$

Since $\langle \cdot \rangle^j$ has to be further percolated in order to eventually apply $[\Box_j I]$, a second K-rule is needed, as explained above. The required rule is (51). Using this rule, the above derivation continues as follows.

$$(51) \quad \Diamond_j (A' \bullet_k B') \rightarrow \Diamond_j A' \bullet_k B'$$

$$\frac{\frac{\langle (\Box_j A \rangle^j \circ_i \Box_j B) \rangle^j \circ_k \Gamma \Rightarrow C}{\langle (\Box_j A \rangle^j \circ_i \Box_j B) \circ_k \Gamma \rangle^j \Rightarrow C} \text{ [(51)]}}{\langle \Box_j A \rangle^j \circ_i \Box_j B \circ_k \Gamma \Rightarrow \Box_j C} \text{ [}\Box_j I\text{]}$$

Suppose MP is *not* applied in line 1. (50) can then be applied only if (51) is used first since only in such case the left component is modally decorated. The result is the sequent in (52).

$$(52) \quad \langle (\Box_j A \circ_k \Gamma) \rangle^j \circ_i \langle \Box_j B \rangle^j \Rightarrow C$$

For the antecedent term, no lexical substitution is possible because there are no lexical items of type $\Box_j A$. Let us finally show how structural rules interact with modal decorations to enforce the use of the former. The general scheme is the following. The percolation mechanism that passes a modal decoration from some substructure to a structure that is of an undecorated designated type has to be construed in such a way that it requires the application of the structural rules. In the above example the use of the rule of permutation creates a context in which the modal decoration

on an np resource can be percolated by the application of two K-rules to the whole complex predicate consisting of VP_1 and VP_2 . Without this percolation no lexical substitution would be possible for the np resource as it requires the modally decorated type $\Box_j A$ and not $\Diamond_j \Box_j A$. Thus, structural rules are used to create contexts which license the percolation of modal decorations, which, in turn, is necessary for lexical substitutions. The above strategy will be key in the derivation of SVCs and CCs which is the topic of the next section.

The discussion in this section has yielded the following strategy for syntactic type assignments in the lexicon in order to enforce the use of structural rules: (i) Modal decorations are used for the syntactic type of both verbs and NPs. Whereas extended verbs are modally decorated by $\Diamond\Box$, NPs are decorated by \Box . For example, a transitive verb in simple sentences or as V_i , $i > 1$, in an SVC or CC (if admissible) is not assigned the syntactic type $np \setminus (s/np)$ but the type $\Diamond\Box((np \setminus_r (s/np)))$. Hence, in addition to the modal decoration, there is a distinction between \cdot_{1l} , the verb-object (left head) mode, and \cdot_{1r} , the subject-verb (right head) mode. If a verb is used as the first verb in an SVC or a CC, one gets $\Diamond\Box((np \setminus_r (s/np)) /_i vp)$, which reflects the fact that there is an additional argument of type vp , (ii) the extended forms of verbs differ at the level of the mode by which the additional argument of type vp combines with the verb, and (iii) the head adjunction modes are \cdot_0 (for CSVCs) and \cdot_2 (for RSVCs and CCs).

THE DERIVATION OF SVCS AND CCS IN EDO

4

The syntactic derivation of CCs and CSVCs with two transitive verbs

4.1

Both in an SVC and a CC with a transitive first verb this verb first combines with a resource of type vp and then with a resource of type np yielding a structure of type vp , which corresponds to the sequent $V_1 VP_2 NP_2$. In order to arrive at the correct word order, which is

V_1 NP_2 VP_2 , the mixed permutation rule MP_1 in (53) is used, with \bullet_i a head adjunction mode.¹³

$$(53) \quad MP1: (A \bullet_{1l} \diamond B) \bullet_i C \rightarrow (A \bullet_i C) \bullet_{1l} \diamond B$$

Note that $MP1$ does not require one of the verbal elements in the verbal cluster to be modally decorated with \diamond . The use of $MP1$ is linked to the use of the K -rule in (54).

$$(54) \quad K^*2(\bullet_{1l}): \diamond(\diamond A \bullet_{1l} B) \rightarrow \diamond A \bullet_{1l} \diamond B$$

This rule requires that the left (verbal) element and the right (nominal) element are both modally decorated. Whereas the decoration of the left component is not percolated, the decoration of the right component is percolated to the whole verbal structure.

Using $MP1$ and $K^*2(\bullet_{1l})$, produces Derivation 1 below:

$$\frac{\frac{[x_1 \Rightarrow \Box(tv/i\text{vp})]^1}{\langle x_1 \rangle \Rightarrow (tv/i\text{vp})} [\Box E] \quad \frac{vp_2 \Rightarrow \text{vp}}{\langle x_1 \rangle \circ_i vp_2 \Rightarrow tv} [/_i E] \quad \frac{np_2 \Rightarrow \Box np}{\langle np_2 \rangle \Rightarrow np} [\Box E]}{\frac{\langle x_1 \rangle \circ_i vp_2 \Rightarrow tv \quad \langle np_2 \rangle \Rightarrow np}{\langle x_1 \rangle \circ_i vp_2 \Rightarrow \text{vp}} [/_i E]} [\Box E]$$

$$\frac{\langle x_1 \rangle \circ_i vp_2 \Rightarrow \text{vp}}{\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle \Rightarrow \text{vp}} [MP1]$$

$$\frac{\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle \Rightarrow \text{vp}}{\langle x_1 \rangle \circ_{1l} np_2 \Rightarrow \text{vp}} [K^*2(\bullet_{1l})]$$

Since the left component is a non-lexical VP , its modal decoration originates from its (nominal) right element and has therefore to be percolated to the whole antecedent structure. This consideration is independent of the exact form of vp_2 . The three possible percolation rules are given in (55).

$$(55) \quad \begin{array}{l} \text{a. } K(\bullet_i): \diamond(A \bullet_i B) \rightarrow \diamond A \bullet_i \diamond B \\ \text{b. } K1(\bullet_i): \diamond(A \bullet_i B) \rightarrow \diamond A \bullet_i B \\ \text{c. } K^*1(\bullet_i): \diamond(A \bullet_i \diamond B) \rightarrow \diamond A \bullet_i \diamond B \end{array}$$

$K(\bullet_i)$ and $K^*1(\bullet_i)$ both require the right component to be modally decorated, too. They differ with respect to the way this decoration is handled. Whereas $K(\bullet_i)$ removes the modal decoration, this is not the

¹³In this and subsequent sections, only the algebraic presentation of structural rules is given. The corresponding inference rule in the natural deduction format can be found in the Appendix.

case for $K^*1(\bullet_i)$. $K1(\bullet_i)$ does not impose any condition on the modal decoration of the right component. It may but need not be modally decorated. Note that $K1(\bullet_i)$ subsumes $K^*1(\bullet_i)$ as a special case.

Deriving the Sequence $V_1 NP_2 V_2 NP_3$ in a CC

4.2

Recall the syntactic structure of a CC with two transitive verbs, exemplified by an example repeated from Section 1.

(56) CC: $NP_1 V_1 NP_2 V_2 NP_3$

Òzó ghá gbè èwé khièn ùhùnmwùn érèn.

Ozo FUT hit goat sell head its

‘Ozo will kill the goat and sell its head.’

Baker and Stewart (1999:3)

In this type of CC, there is an overt NP after V_2 , which is in addition not required to be coreferential with the direct object of V_1 (= NP_2). In derivation 1, vp_2 is therefore a structure of the form $\langle\langle x_2 \rangle \circ_{1l} np_3\rangle$ of type vp , i.e. a non-lexical VP. Consequently, its modal decoration originates from the NP argument and therefore has to be passed to the whole antecedent structure, i.e. to the sequence corresponding to the complex VP = $V_1 NP_2 V_2 NP_3$. Thus, both components of \circ_i are structures of the form $\langle\langle x \rangle \circ_{1l} np\rangle$, corresponding to a non-lexical VP. The required K-rule therefore is (55a), which distributes \diamond over both components. Setting the head adjunction mode to \cdot_2 , one gets (57).

(57) $K(\bullet_2): \diamond(A \bullet_2 B) \rightarrow \diamond A \bullet_2 \diamond B$

Given $K(\bullet_2)$ and setting $vp_2 = \langle\langle x_2 \rangle \circ_{1l} np_3\rangle$ and $\circ_i = \circ_2$, the Derivation 1 from above continues as follows.

$$\frac{\langle\langle x_1 \rangle \circ_{1l} np_2\rangle \circ_2 \langle\langle x_2 \rangle \circ_{1l} np_3\rangle \Rightarrow vp}{\langle\langle x_1 \rangle \circ_{1l} np_2\rangle \circ_2 \langle\langle x_2 \rangle \circ_{1l} np_3\rangle \Rightarrow vp} [K(\bullet_2)]$$

So far, we have shown how the assumed structural rules enable deriving a sequent of type vp with the correct word order corresponding to the complex VP = $V_1 NP_2 V_2 NP_3$. It remains to show that they also *enforce* it. Suppose in line 4 in Derivation 1 from above, repeated below with the necessary substitution, the rule MP_1 is not applied.

$$4. \quad (\langle x_1 \rangle \circ_2 (\langle x_2 \rangle \circ_{1l} np_3)) \circ_{1l} \langle np_2 \rangle \Rightarrow vp$$

The structure in the antecedent is of the form $\Gamma \circ_{1l} \langle \Delta \rangle$. Since the structural operator on the right component has to be percolated to the antecedent term, rule $K^*2(\bullet_{1l})$ has to be applied. This is possible only if rule $K(\bullet_2)$ has been applied to the left component since $K^*2(\bullet_{1l})$ requires that the left component be modally decorated. Application of this rule yields line 5*.

$$5^*. \quad (\langle x_1 \circ_2 (\langle x_2 \rangle \circ_{1l} np_3) \rangle) \circ_{1l} \langle np_2 \rangle \Rightarrow vp$$

This step is fatal because the modal decoration of the left component is percolated by $K(\bullet_2)$. Consequently, since x_1 is of type $\Box(tv/_2vp)$, the sequent requires a lexical element that is of that type. But there are no such lexical entries, transitive verbs being of a type that is modally decorated with $\Diamond\Box$: $\Diamond\Box tv$ or $\Diamond\Box(tv/_i vp)$. As a result, the sequent in line 5* does not admit a substitution of lexical elements. To put it differently, removing the decoration of the left component, it is no longer possible to apply $[\Diamond E]$ at a later stage, using the lexical axiom $v_1 \Rightarrow \Diamond\Box(tv/_i vp)$.¹⁴

Let us analyze the success and the failure in more detail. $K(\bullet_2)$ requires the left component of \bullet_i , $i = 0$ or $i = 2$, to be \Diamond -decorated. In the intended case, in which MPI is applied, this left component does not correspond to the extended verb ($=V_1$) but to the VP built in terms of this verb. Assuming that $K^*2(\bullet_{1l})$ has been applied, this component is of the form $\langle \langle \Gamma \rangle \circ_{1l} \Delta \rangle$ with $\langle \Gamma \rangle$ corresponding to V_1 and Δ corresponding to the object argument of V_1 . In this case the

¹⁴One has the derived rule below, which is the left rule for \Diamond in a Gentzen sequent presentation

$$(*) \quad \frac{\Gamma[\langle A \rangle] \Rightarrow C}{\Gamma[\langle \Diamond A \rangle] \Rightarrow C}$$

Therefore, in a non-sugared presentation one has (with $\alpha = tv$ or $\alpha = (tv/_i vp)$)

$$(**) \quad \frac{\Gamma[\langle \Box \alpha \rangle] \Rightarrow C}{\Gamma[\langle \Diamond \Box \alpha \rangle] \Rightarrow C}$$

Since there are lexical items of type $\Diamond\Box\alpha$, they can be substituted for an occurrence of this categorial formula in Γ . After removing the modal decoration, the step (**) is no longer possible.

outer \diamond -decoration should be passed to the whole structure since it originated from the decoration of the NP argument which should be percolated to the whole structure.

By contrast, in the derivation yielding the incorrect word order, the order in which $K(\bullet_2)$ and $K^*2(\bullet_{1l})$ are applied is reversed. This is the case because $K^*2(\bullet_{1l})$ requires the left component to be modally decorated. Contrary to the intended case, the left component of the verbal cluster composed by \circ_i is a resource corresponding to V_1 and *not* to the VP built from it. This is a simple consequence of the fact that permutation has not yet been applied so that the linear order corresponds to the order in which the arguments are discharged. Since $K(\bullet_2)$ removes the decoration of the left component, the result is linguistically ill-formed because it requires a resource of type $\square(tv/{}_i vp)$. However, there happen to be no lexical entries meeting this condition.

The above argument only requires a percolation rule involving a head adjunction mode to remove the decoration of the left component. As was shown above in the preceding section, this condition is satisfied by all possible percolation rules. Thus, the argument equally applies if instead of $K(\bullet_2)$ $K1(\bullet_i)$ or $K^*1(\bullet_i)$ is used. The failure of a derivation in which the mixed permutation rule is not applied becomes even more apparent in the non-sugared presentation.

$$\frac{\frac{\frac{\square(tv/{}_i vp) \Rightarrow \square(tv/{}_i vp)}{\langle \square(tv/{}_i vp) \rangle \Rightarrow (tv/{}_i vp)} \text{ [}\square\text{E]}}{\langle \square(tv/{}_i vp) \rangle \circ_i vp_2 \Rightarrow tv} \text{ [}/\text{E]}}{\frac{\frac{\frac{vp_2 \Rightarrow vp}{np_2 \Rightarrow \square np} \text{ [}\square\text{E]}}{\langle np_2 \rangle \Rightarrow np} \text{ [}/\text{E]}}{\langle \langle \square(tv/{}_i vp) \rangle \circ_i vp_2 \rangle \circ_{1l} \langle np_2 \rangle \Rightarrow vp} \text{ [K-rule for } \bullet_i\text{]}}{\frac{\langle \square(tv/{}_i vp) \rangle \circ_i vp_2 \rangle \circ_{1l} \langle np_2 \rangle \Rightarrow vp} \text{ [K}^*\text{2}(\bullet_{1l})\text{]}}{\langle \langle \square(tv/{}_i vp) \rangle \circ_i vp_2 \rangle \circ_{1l} np_2 \rangle \Rightarrow vp}$$

In addition, application of $K^*2(\bullet_{1l})$ does *not* remove the modal decoration from the verbal cluster, as the last line 6 shows. As a consequence, application of rule $[\diamond E]$ to this line requires a verbal cluster $(x_1 \circ_i vp_2)$ to be of type $\diamond \square tv$, i.e. $(v_1 \circ_i vp_2) \Rightarrow \diamond \square tv$, with $x_1 \Rightarrow \square(tv/{}_i vp)$, which is not derivable.

The above discussion has shown that a percolation rule involving a head adjunction mode has to be applied *after* the rule $K^*2(\bullet_{1l})$ has been applied in order to work correctly. Consequently, the order

in which the rules are applied matters. This order is sensitive to the application of the rule of permutation MPI. If it is applied, the order in which the K-rules are applied is the correct one, otherwise not. To put it differently, the correct order requires a structure of the form (58a) and not a structure of the form (58b). The effect of MPI is just to transform (58b) into (58a).

- (58) a. $(\langle \Gamma \rangle \circ_{1l} \langle \Delta \rangle) \circ_i \Delta'$
 b. $(\langle \Gamma \rangle \circ_i \Delta') \circ_{1l} \langle \Delta \rangle$

Applying $K^*2(\bullet_{1l})$ and one of the percolation rules for the head adjunction modes in the wrong order always yields sequents that do not admit lexical substitutions for the terms in the antecedent.

The problem of getting the correct order of rule applications can be solved by distinguishing two different kinds of phrasal structures of type vp: $(\langle x_1 \rangle \circ_{1l} \langle np \rangle)$ and $((\langle x_1 \rangle \circ_i vp) \circ_{1l} \langle np \rangle)$. Only the first is linguistically admissible, in which the left component of \circ_{1l} is *not* a verbal cluster consisting of two verbs. The task, therefore, is reduced to distinguishing such clusters from simple verbs in the contexts of a left-headed phrasal structure. A first key in achieving this consists in modally decorating transitive verbs in the lexicon in such a way that first they enter a derivation as structures modally decorated with \diamond (or $\langle \cdot \rangle$) and second this decoration must not be percolated until a structure of type vp is built up (i.e. until application of rule $K^*2(\bullet_{1l})$). This is achieved by assigning transitive verbs the types $\diamond \square tv$ and $\diamond \square (tv/i vp)$, $i = 0$ or $i = 2$. The second key consists in letting rule $K^*2(\bullet_{1l})$ be sensitive to this modal decoration in the sense that it is explicitly checked whether the component is modally decorated. Since verbal clusters are not lexical in Edo, one arrives at a structure of the form required by rule $K^*2(\bullet_{1l})$ only if a percolation rule for a head adjunction mode is applied. But, and this is the third key, these rules remove the modal decoration of the left component of the verbal cluster, i.e. of the extended verb, so that it is no longer possible to find a lexical substitution.

The modal decoration of transitive verbs, therefore, functions as a domain modality. In the context of structures composing a verbal element and a direct object it admits to distinguish simple transitive verbs from verbal clusters both of which can be composed

with an np resource by \circ_{1l} due to the mixed permutation rule MP1. Whereas the former are modally decorated without application of a percolation rule, the latter are modally decorated only if such a rule is applied. Thus, rule $K^*2(\bullet_{1l})$ can be said to require *lexical* verbal heads.

The failure that results if MP1 is not applied can also be shown by trying to parse an expression of type vp with the incorrect word order.

$$\begin{array}{c}
 \text{fail} \\
 \hline
 \frac{\langle (\langle \Box(\text{tv}/_2\text{vp}) \rangle \circ_2 (\langle \Box\text{tv} \rangle \circ_{1l} \Box\text{np})) \circ_{1l} \Box\text{np} \rangle \Rightarrow \text{vp}}{\langle (\langle \Box(\text{tv}/_2\text{vp}) \rangle \circ_2 (\langle \Box\text{tv} \rangle \circ_{1l} \Box\text{np})) \circ_{1l} \Box\text{np} \rangle \Rightarrow \Box\text{vp}} \quad [\Box] \\
 \hline
 \langle \Diamond \Box(\text{tv}/_2\text{vp}) \rangle \circ_2 \langle \Diamond \Box\text{tv} \circ_{1l} \Box\text{np} \rangle \circ_{1l} \Box\text{np} \Rightarrow \Box\text{vp} \quad [*]
 \end{array}$$

The derivation already stops at the third line, which is of the form $\langle \Gamma \circ_{1l} \Delta \rangle \Rightarrow \text{vp}$, because application of $K^*2(\bullet_{1l})$ requires the left component to be modally decorated. Yet it is only possible to get $\langle (\langle \Box(\text{tv}/_2\text{vp}) \rangle \circ_2 (\langle \Box\text{tv} \rangle \circ_{1l} \Box\text{np})) \rangle$ since this component is *not* a lexical verbal head.

Deriving the sequence $V_1 NP_2 V_2$ in a CSVC

4.3

In contrast to a CC, the object arguments of V_1 and V_2 are identified with each other in a CSVC and the direct object of V_2 cannot be overtly realized, either as an NP or as a pronoun which is coreferential with NP_2 (= the DO of V_1). Below, we repeat an example from Section 1.

(59) CSVC: $NP_1 V_1 NP_2 V_2$

Òzó ghá gbè èwé khièn.

Ozo FUT hit goat sell

‘Ozo will kill the goat and sell it.’

Baker and Stewart (1999:3)

If both verbs in a CSVC are transitive and the additional argument of the extended first verb is of type vp, one gets Derivation 2 below assuming the head adjunction mode to be \circ_0 :

$$\begin{array}{c}
 \frac{[x_1 \Rightarrow \Box(tv/_0vp)]^2}{\langle x_1 \rangle \Rightarrow tv/_0vp} \quad [\Box E] \quad \frac{[x_2 \Rightarrow \Box tv]^1}{\langle x_2 \rangle \Rightarrow tv} \quad [\Box E] \quad \frac{np_2 \Rightarrow \Box np}{\langle np_2 \rangle \Rightarrow np} \quad [\Box E]}{\frac{\langle x_2 \rangle \circ_{1l} \langle np_2 \rangle \Rightarrow vp}{\langle x_1 \rangle \circ_0 (\langle x_2 \rangle \circ_{1l} \langle np_2 \rangle) \Rightarrow tv} \quad [/_0E]} \quad \frac{np_2 \Rightarrow \Box np}{\langle np_2 \rangle \Rightarrow np} \quad [\Box E]}{\frac{((\langle x_1 \rangle \circ_0 (\langle x_2 \rangle \circ_{1l} \langle np_2 \rangle)) \circ_{1l} \langle np_2 \rangle) \Rightarrow vp}{(\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle) \circ_0 (\langle x_2 \rangle \circ_{1l} \langle np_2 \rangle) \Rightarrow vp} \quad [MP1]} \quad [/_1E]
 \end{array}$$

Up to that point the derivation is parallel to that for a CC with two transitive verbs, except that the head adjunction modes are assumed to be different and that the np resource np_2 has been used twice. This second difference reflects the fact that in a CSVC the DO are identified and that the DO of V_2 cannot be overtly realized. Consequently, in a CSVC, the np resource corresponding to the shared DO has to be used twice if it is assumed that the additional argument by which the subcategorization frame of V_1 is extended is of type vp. It is used both as the object argument of V_1 and as the object argument of V_2 . From what has been said it follows that at line 6 a rule of Mixed Contraction has to be applied. In the present context, it takes the form (60).

$$(60) \quad MC: (A \bullet_0 B) \bullet_{1l} \diamond C \rightarrow (A \bullet_{1l} \diamond C) \bullet_0 (B \bullet_{1l} \diamond C)$$

Applying MC to line 6 in Derivation 2 yields line 7.

$$7. \quad (\langle x_1 \rangle \circ_0 \langle x_2 \rangle) \circ_{1l} \langle np_2 \rangle \Rightarrow vp$$

After the rule of mixed contraction has been applied, the np resource must again be infixed in the verbal cluster, using the rule MP1 of mixed permutation. This gives line 8.

$$8. \quad (\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle) \circ_0 \langle x_2 \rangle \Rightarrow vp$$

Comparing this line with line 6 in Derivation 1 of a CC, one notices that in a CSVC vp_2 ultimately is only V_2 since the object argument has been elided due to the application of the rule of mixed contraction. Thus, it is a structure of the form $\langle x \rangle$ with x of type $\Box tv$. The modal decoration of the right component of a \circ_0 -structure must therefore not be percolated. The appropriate percolation rule for \bullet_0 is therefore (61), which distributes \diamond only over the left component.

$$(61) \quad K1(\bullet_0): \diamond(A \bullet_0 B) \rightarrow \diamond A \bullet_0 B$$

Applying $K1(\bullet_0)$ to line 8 yields line 9.

$$9. \quad \langle (\langle x_1 \rangle \circ_{1l} np_2) \circ_0 \langle x_2 \rangle \rangle \Rightarrow vp$$

In the derivation of a CSVC the rule MP1 is used twice. In both cases an np resource is infixed in a verbal cluster. In the first application this verbal cluster has the form $(\langle x_1 \rangle \circ_0 (\langle x_2 \rangle \circ_{1l} \langle np_2 \rangle))$. In this situation application of MP1 is enforced because otherwise the only way to proceed consists in first applying $K^*2(\bullet_{1l})$ to $(\langle x_2 \rangle \circ_{1l} \langle np_2 \rangle)$ and then $K1(\bullet_0)$ to $(\langle x_1 \rangle \circ_0 \langle \langle x_2 \rangle \circ_{1l} np_2 \rangle)$, which percolates the structural operator of the left but not that of the right component. As a result, no lexical substitution is possible because the undecorated x_1 is of type $\square(tv/{}_0vp)$ and there are no extended verbs of this type. The problem is that $K1(\bullet_0)$ works correctly only if the verbal cluster consists of a left component that corresponds to a non-lexical VP, i.e. it is of the form $(\langle x \rangle \circ_{1l} np)$, whereas the right component is a verbal element, i.e. it is of the form $\langle x' \rangle$ in the case of a CSVC. One arrives at such a structure only by applying MP1 (and, in addition, MC). The second application of the rule MP1 occurs after contraction so that the right component of the verbal cluster is no longer of the form $(\langle x_2 \rangle \circ_{1l} \langle np_2 \rangle)$ but of the form $\langle x_2 \rangle$. This application is enforced too for the same reasons the previous applications of this rule have been enforced: a verbal cluster is composed with a nominal element to its right.

If in line 6 of Derivation 2 rule MC is not applied, applying $K^*2(\bullet_{1l})$ to both components of the antecedent term yields substructures of the form $(\langle x \rangle \circ_{1l} np)$. Since $K1(\bullet_0)$ only removes the modal decoration of the left component of a structure composed by \circ_0 , the modal decoration of the right component is left intact. Application of $[\diamond E]$ to this component is not possible because this requires the derivability of the sequent $(\square tv \circ_0 \square np) \Rightarrow \diamond \square vp$. Even if this sequent were derivable, its antecedent term does not admit substituting lexical items for the left component since there are no lexical items of type $\square tv$.

Since both in a CSVC and in a CC the sequent in (62) below is derived, it is necessary to distinguish two different kinds of head adjunction modes. With respect to this sequent, the two types of constructions are structurally indistinguishable. In order to enforce the

difference that results beginning from that sequent, principally due to the application of the rule MC in the CSV, two head adjunction modes must be used for which different structural rules apply.

$$(62) \quad (\langle x_1 \rangle \circ_{1l} \langle np \rangle) \circ_i (\langle x_2 \rangle \circ_{1l} \langle np \rangle) \Rightarrow vp$$

4.4

*Deriving the sequence NP₁ VP:
a structural rule for the subject argument*

The rules in (54) and (57) must be supplemented with a corresponding rule for the composition of the subject argument with the VP. From the discussion so far it follows that the sequence V₁ NP₂ V₂ (NP₃) in a CSV or a CC corresponds to a sequent of the form $\langle \Gamma \rangle \Rightarrow vp$. Since the external argument corresponds to a sequent of the form $\langle np_1 \rangle \Rightarrow np$, composing the two resources requires the following percolation rule for \bullet_{1r} , which is the composition mode for right-headed head-complement structures (subject-verb relation).

$$(63) \quad K(\bullet_{1r}): \diamond(A \bullet_{1r} B) \rightarrow \diamond A \bullet_{1r} \diamond B$$

The justification of $K(\bullet_{1r})$ runs as follows. First, the \diamond -decoration of an np resource has to be percolated. Second, the \diamond -decoration of any non-minimal verbal projection of a transitive verb has to be percolated since it originates from the decoration of an NP complement.¹⁵ The relevant derivation is given below.

$$\frac{\frac{\frac{np_1 \Rightarrow \square np}{\langle np_1 \rangle \Rightarrow np} [\square E] \quad \langle vp \rangle \Rightarrow vp}{\langle np_1 \rangle \circ_{1r} \langle vp \rangle \Rightarrow s} [\wedge_{1r} E]}{\frac{\langle np_1 \rangle \circ_{1r} \langle vp \rangle \Rightarrow s}{\langle np_1 \circ_{1r} vp \rangle \Rightarrow s} [K(\bullet_{1r})]} [\square I]}{np_1 \circ_{1r} vp \Rightarrow \square s} [\square I]$$

Given the K-rule for the subject argument, the complete derivations for a CC and a CSV with two transitive verbs are given below. We start with a CC. The derivation is displayed on page 377.

Since we finally derived objects of syntactic type $\square s$, we will also provide information about the semantics. For the sake of readability,

¹⁵This argument also holds for verbal VPs, i.e. a VP projected by an intransitive verb; see Section 4.7 for details.

CC (two transitive verbs):

$$\begin{array}{c}
 \frac{[x_1 \Rightarrow \Box(tv/{}_2vp)]^2}{\langle x_1 \rangle \Rightarrow tv/{}_2vp} \quad \frac{[x_2 \Rightarrow \Box tv]^1}{\langle x_2 \rangle \Rightarrow tv} \quad \frac{np_3 \Rightarrow \Box np}{\langle np_3 \rangle \Rightarrow np} \quad \frac{np_2 \Rightarrow \Box np}{\langle np_2 \rangle \Rightarrow np} \\
 \frac{\frac{\frac{\langle x_1 \rangle \circ_2 (\langle x_2 \rangle \circ_{1l} \langle np_3 \rangle) \Rightarrow tv}{((\langle x_1 \rangle \circ_2 (\langle x_2 \rangle \circ_{1l} \langle np_3 \rangle)) \circ_{1l} \langle np_2 \rangle) \Rightarrow vp} \quad \frac{[MP1]}{((\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle) \circ_2 (\langle x_2 \rangle \circ_{1l} \langle np_3 \rangle)) \Rightarrow vp} \quad \frac{[K^{**2}(\bullet_{1l})]}{\text{on both comp.}}}{\langle x_1 \rangle \circ_{1l} np_2 \circ_2 \langle x_2 \rangle \circ_{1l} np_3 \Rightarrow vp} \quad \frac{[K(\bullet_2)]}{[_1rE]} \\
 \frac{np_1 \Rightarrow \Box np}{\langle np_1 \rangle \Rightarrow np} \quad \frac{((\langle x_1 \rangle \circ_{1r} ((\langle x_1 \rangle \circ_{1l} np_2) \circ_2 (\langle x_2 \rangle \circ_{1l} np_3))) \Rightarrow s}{\langle np_1 \circ_{1r} ((\langle x_1 \rangle \circ_{1l} np_2) \circ_2 (\langle x_2 \rangle \circ_{1l} np_3)) \rangle \Rightarrow s} \quad \frac{[K(\bullet_{1r})]}{[\Box I]} \\
 \frac{np_1 \circ_{1r} ((\langle x_1 \rangle \circ_{1l} np_2) \circ_2 (\langle x_2 \rangle \circ_{1l} np_3)) \Rightarrow \Box s}{np_1 \circ_{1r} ((\langle x_1 \rangle \circ_{1l} np_2) \circ_2 (\langle x_2 \rangle \circ_{1l} np_3)) \Rightarrow \Box s} \quad \frac{v_1 \Rightarrow \Diamond \Box (tv/{}_2vp)}{v_2 \Rightarrow \Diamond \Box tv} \\
 \frac{np_1 \circ_{1r} ((v_1 \circ_{1l} np_2) \circ_2 (v_2 \circ_{1l} np_3)) \Rightarrow \Box s}{np_1 \circ_{1r} ((v_1 \circ_{1l} np_2) \circ_2 (\langle x_2 \rangle \circ_{1l} np_3)) \Rightarrow \Box s} \quad \frac{[\Diamond E]^2}{[\Diamond E]^1}
 \end{array}$$

we will not annotate the syntactic proof tree with semantic terms. Instead, we follow a common practice and only give the semantic term at the end of a derivation together with an example from Section 1. We translate proper names, common nouns and mass nouns as expressions of type e . There are two reasons for this. Since we do not examine quantification in this article, we choose the most simple translation. On the empirical side, one has that ‘bare’ common nouns in Edo are standardly interpreted as singular definite expression ‘the cn ’. We assume this standard interpretation also for mass nouns and use the iota-operator: $cn \rightarrow \iota z.cn(z)$ and the same for mass nouns.¹⁶ The interpretation of λ terms is given in the Appendix.

Recall that in a CC there is no constraint that the direct objects have to be shared. For (64) in which the direct objects are different, one gets (65a) as derivational semantics. When substituting the lexical semantics into this derivational semantics using the meaning of ‘gboo’ (plant) in (65c) one gets (65b).¹⁷ Note that for V_1 ‘gboo’ (plant) the extended form is used at the syntactic level and, therefore, the complex meaning in (65c).¹⁸

- (64) Òzó_k gbòó ìvìn b̀l̀ó ókà.
 Ozo plant coconut peel corn
 ‘Ozo planted coconut and peeled the corn.’
 Stewart (2001:65)

- (65) a. $((x_{v_1}(x_{v_2}x_{np_3}))x_{np_2})x_{np_1}$.
 b. $\lambda e.\exists e_1.\exists e_2[e = e_1 \wedge plant(e_1) \wedge peel(e_2) \wedge actor(e_1) = ozo \wedge theme(e_1) = \iota w.coconut(w) \wedge actor(e_1) = actor(e_2) \wedge theme(e_2) = \iota z.corn(z) \wedge e_1 \preceq e_2]$.
 c. $\lambda VP_2.\lambda y.\lambda x.\lambda e.\exists e_1.\exists e_2[e = e_1 \wedge plant(e_1) \wedge VP_2(x)(e_2) \wedge actor(e_1) = x = first(e_2) \wedge e_1 \preceq e_2]$.

¹⁶Though the translation contains a term of type $\langle e, t \rangle$, i.e. cn , this term is not used as the translation of ‘ cn ’.

¹⁷In (65b) we already applied simplifications related to thematic roles using equational reasoning. We did not apply the simplification $e = e_1$ in order to highlight the similarities and differences to SVCs.

¹⁸The original example in Stewart (2001) has an overt subject pronoun which is left out in (64).

If the direct objects are identified, the direct object of V_2 is realized by a pronoun. A proper analysis of CCs in which the direct objects are shared requires an interpretation of pronouns in a dynamic semantics. Since such an analysis is beyond the scope of this article, we make the following assumption. Similarly to Dynamic Predicate Logic and Compositional Discourse Representation Theory, it is assumed that anaphora-antecedent relationships are represented at the level of logical form in the form of preindexation so that the antecedent of a pronoun is known.¹⁹ Using (65a) and (65c), one gets (67) for (66).

- (66) Òzó_k lé ízè_j Ó_k rrí órè_j.
 Ozo cook rice he eat it
 ‘Ozo cooked rice and he ate it.’
 Stewart (2001:64)

- (67) $\lambda e. \exists e_1. \exists e_2 [e = e_1 \wedge \text{plant}(e_1) \wedge \text{peel}(e_2) \wedge \text{actor}(e_1) = \text{ozo} \wedge \text{theme}(e_1) = \iota w. \text{rice}(w) \wedge \text{actor}(e_1) = \text{actor}(e_2) \wedge \text{theme}(e_1) = \text{theme}(e_2) \wedge e_1 \preceq e_2]$.

Next we turn to a CSV. The derivation is displayed on page 380.

For an illustration of the semantic derivation of a CSV, we will use the example in (68). The derivational semantics is given in (69a). Applying (69c) to the representation of VP_2 and the two arguments of V_1 yields (69b). Note that also in this case the extended verb form is used for V_1 and hence the (complex) meaning.

- (68) Òzó lé èvbàré ré.
 Ozo cook food eat
 ‘Ozo cooked food and ate it.’
 Stewart (2001:60)

- (69) a. $((x_{v_1}(x_{v_2}x_{np_2}))x_{np_2})x_{np_1}$.
 b. $\lambda e. \exists e_1. \exists e_2 [e = e_1 \sqcup e_2 \wedge \text{cook}(e_1) \wedge \text{eat}(e_2) \wedge \text{actor}(e_1) = \text{ozo} \wedge \text{theme}(e_1) = \iota z. \text{food}(z) \wedge \text{actor}(e_1) = \text{actor}(e_2) \wedge \text{theme}(e_1) = \text{theme}(e_2) \wedge e_1 \preceq e_2 \wedge \Box_{\text{ozo}}(\text{occur}(e_1) \rightarrow \text{occur}(e_2))]$.

¹⁹ See Jäger (2005) for an analysis of pronouns in TLG.

- c. $\lambda VP_2. \lambda y. \lambda x. \lambda e. \exists e_1. \exists e_2 [e = e_1 \sqcup e_2 \wedge \text{cook}(e_1) \wedge VP_2(x)(e_2) \wedge \text{actor}(e_1) = x = \text{first}(e_2) \wedge \text{theme}(e_1) = y = \text{second}(e_2) \wedge e_1 \preceq e_2 \wedge \Box_x(\text{occur}(e_1) \rightarrow \text{occur}(e_2))]$.

The derivation of simple sentences with transitive verbs

4.5

So far, CSVCs and CCs in which both verbs are transitive have been considered. In order to show the theory to be successful it is necessary to be able to also derive simple sentences with transitive verbs. The derivation is given below.

Simple Sentence (transitive verb):

$$\begin{array}{c}
 \frac{\text{np}_1 \Rightarrow \Box \text{np}}{\langle \text{np}_1 \rangle \Rightarrow \text{np}} \text{[\Box E]} \quad \frac{[x \Rightarrow \Box \text{tv}]^1}{\langle x \rangle \Rightarrow \text{tv}} \text{[\Box E]} \quad \frac{\text{np}_2 \Rightarrow \Box \text{np}}{\langle \text{np}_2 \rangle \Rightarrow \text{np}} \text{[\Box E]} \\
 \frac{\langle x \rangle \circ_{1l} \langle \text{np}_2 \rangle \Rightarrow \text{vp}}{\langle \text{np}_1 \rangle \circ_{1r} (\langle x \rangle \circ_{1l} \langle \text{np}_2 \rangle) \Rightarrow \text{s}} \text{[\Box E]} \quad \frac{\langle \text{np}_1 \rangle \circ_{1r} (\langle x \rangle \circ_{1l} \langle \text{np}_2 \rangle) \Rightarrow \text{s}}{\langle \text{np}_1 \rangle \circ_{1r} \langle \langle x \rangle \circ_{1l} \text{np}_2 \rangle \Rightarrow \text{s}} \text{[K*2(\bullet_{1l})]} \\
 \frac{\langle \text{np}_1 \rangle \circ_{1r} \langle \langle x \rangle \circ_{1l} \text{np}_2 \rangle \Rightarrow \text{s}}{\langle \text{np}_1 \rangle \circ_{1r} (\langle x \rangle \circ_{1l} \text{np}_2) \Rightarrow \text{s}} \text{[K(\bullet_{1r})]} \quad v \Rightarrow \Diamond \Box \text{tv} \text{[\Diamond E]}^1 \\
 \frac{\langle \text{np}_1 \rangle \circ_{1r} (v \circ_{1l} \text{np}_2) \Rightarrow \text{s}}{(\text{np}_1 \circ_{1r} (v \circ_{1l} \text{np}_2)) \Rightarrow \Box \text{s}} \text{[\Box I]}
 \end{array}$$

Since in a simple sentence with a transitive verb the latter is not extended, it is of type $\Diamond \Box \text{tv}$ rather than of type $\Diamond \Box (\text{tv}/_i \text{vp})$. Similarly to a CSVC and a CC, the derivation starts with hypothetically assuming a resource of type $\Box \text{tv}$, which gets eventually discharged using $v \Rightarrow \Diamond \Box (\text{tv}/_i \text{vp})$ and $[\Diamond E]$. After composing x with np_2 to form a vp , $K^*2(\bullet_{1l})$ is applied, percolating the \Diamond -decoration of the right but not that of the left component. The result is the structure $\langle \langle x \rangle \circ_{1l} \text{np}_2 \rangle$. This structure is next composed with the structure corresponding to the subject argument. Applying $K(\bullet_{1r})$ to the resulting structure, percolates both \Diamond -decorations, yielding the structure $\langle \text{np}_1 \circ_{1r} (\langle x \rangle \circ_{1l} \text{np}_2) \rangle$ of type s . Next, the hypothetical assumption is discharged. Finally, application of $[\Box I]$, gives the last line of the derivation. Thus, this argument actually reproduces that for the corresponding substructures in a CSVC or CC. The semantic level is illustrated with (70).

- (70) Òzó lé èvbàré.
 Ozo cook food
 ‘Ozo cooked the food.’
 Stewart (2001:44)

- (71) a. $((x_v x_{np_2}) x_{np_1})$.
 b. $\lambda e.[\text{cook}(e) \wedge \text{actor}(e) = \text{ozo} \wedge \text{theme}(e) = \iota z.\text{food}(z)]$.

4.6 *The derivation of CCs and simple sentences with intransitive verbs*

For a CC and an RSVC, both verbs can be intransitive. From the possibility that intransitive verbs can occur as the first verb in multiverb sequences it follows that they too can have an extended subcategorization frame. This does not mean, however, that the modal decoration for intransitive verbs, either extended or not, is the same as that for transitive verbs. The choice of a modal decoration is, of course, already restricted by the rules that have been assumed for the derivation of CSVCS and CCs with two transitive verbs. In particular, the two structural rules distributing the unary connective \diamond across compositions of a verb with one of its default subcategorized arguments (i.e. either the subject or the object argument) are required to hold for RSVCs and CCs with intransitive verbs, too. This constraint already excludes a modal decoration of the form $\diamond\Box$ that has been used for transitive verbs in the lexicon. In a simple sentence with an intransitive verb the VP usually consists only of the verb since there is no argument to the right of the verb with which it combines first. Consequently, only $K(\bullet_{1r})$ applies. Assuming intransitive verbs to be of type $\diamond\Box\text{vp}$, one gets the derivation below.

$$\frac{\frac{\frac{\text{np}_1 \Rightarrow \Box\text{np}}{\langle \text{np}_1 \rangle \Rightarrow \text{np}} \quad [\Box\text{E}]}{\langle \text{np}_1 \rangle \circ_{1r} \langle x \rangle \Rightarrow \text{S}} \quad [\text{K}(\bullet_{1r})]}{\langle \text{np}_1 \circ_{1r} x \rangle \Rightarrow \text{S}} \quad [\Box\text{I}]}{\frac{[\text{K}(\bullet_{1r})]}{\langle x \rangle \Rightarrow \text{vp}} \quad [\setminus_{1r}\text{E}]}{[x \Rightarrow \Box\text{vp}]^1} \quad [\Box\text{E}]}{\langle x \rangle \Rightarrow \text{vp}} \quad [\setminus_{1r}\text{E}]}$$

Since the vp resource is of the form $\langle \Gamma \rangle$, its decoration is percolated by the application of $K(\bullet_{1r})$. But this means that it is no longer possible to apply the lexical axiom $v \Rightarrow \diamond\Box\text{vp}$ to x , using the rule $[\diamond\text{E}]$ in order to discharge the hypothetical assumption and get a possible lexical substitution for the final antecedent term. The problem is that $K(\bullet_{1r})$ was introduced in the first place for VPs that are built from a vp and an np resource, i.e. for non-lexical VPs. In this case, as has been

shown in the preceding section, the \diamond -decoration of the right component originates from the np resource and should therefore be passed to the whole structure of type s in order to license application of the $[\square I]$ rule.

The failure of the above derivation already shows a possible solution. An intransitive verb is assigned the type $\square vp$ in the lexicon. One then gets the following derivation, which poses no problem.

Simple Sentence (intransitive verb):

$$\frac{\frac{\frac{np_1 \Rightarrow \square np}{\langle np_1 \rangle \Rightarrow np} \quad [\square E] \quad \frac{v \Rightarrow \square vp}{\langle v \rangle \Rightarrow vp} \quad [\square E]}{\langle np_1 \rangle \circ_{1r} \langle v \rangle \Rightarrow s} \quad [\wedge_{1r} E]}{\frac{\langle np_1 \rangle \circ_{1r} \langle v \rangle \Rightarrow s}{\langle np_1 \rangle \circ_{1r} v \Rightarrow s} \quad [K(\bullet_{1r})]} \quad [\square I]}{\langle np_1 \rangle \circ_{1r} v \Rightarrow \square s}$$

We illustrate the semantic derivation with (72).

- (72) Òzó dé.
Ozo fall
'Ozo fell.'
Stewart (2001:87)

- (73) a. $x_v x_{np_2} x_{np_1}$.
b. $\lambda e.[fall(e) \wedge theme(e) = ozo]$.

For a CC with a transitive first and an intransitive second verb one gets the derivation presented below.

CC (transitive and intransitive verb):

$$\frac{\frac{\frac{\frac{[x_1 \Rightarrow \square (tv/2vp)]^1}{\langle x_1 \rangle \Rightarrow tv/2vp} \quad [\square E] \quad \frac{\frac{v_2 \Rightarrow \square vp}{\langle v_2 \rangle \Rightarrow vp} \quad [\square E]}{\langle x_1 \rangle \circ_2 \langle v_2 \rangle \Rightarrow tv} \quad [/_2 E]}{\langle x_1 \rangle \circ_2 \langle v_2 \rangle \Rightarrow tv} \quad [/_1 E]} \quad \frac{\frac{np_2 \Rightarrow \square np}{\langle np_2 \rangle \Rightarrow np} \quad [\square E]}{\langle x_1 \rangle \circ_2 \langle v_2 \rangle \circ_{1l} \langle np_2 \rangle \Rightarrow vp} \quad [MP1]}{\frac{\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle \circ_2 \langle v_2 \rangle \Rightarrow vp}{\langle \langle x_1 \rangle \circ_{1l} np_2 \rangle \circ_2 \langle v_2 \rangle \Rightarrow vp} \quad [K^*2(\bullet_{1l})]} \quad [K(\bullet_2)]}{\frac{\langle \langle x_1 \rangle \circ_{1l} np_2 \rangle \circ_2 \langle v_2 \rangle \Rightarrow vp}{\langle np_1 \rangle \circ_{1r} \langle \langle x_1 \rangle \circ_{1l} np_2 \rangle \circ_2 v_2 \rangle \Rightarrow s} \quad [\wedge_{1r} E]} \quad [K(\bullet_{1r})]}{\frac{\langle np_1 \rangle \circ_{1r} \langle \langle x_1 \rangle \circ_{1l} np_2 \rangle \circ_2 v_2 \rangle \Rightarrow s}{\langle np_1 \rangle \circ_{1r} ((\langle x_1 \rangle \circ_{1l} np_2) \circ_2 v_2)} \Rightarrow s} \quad [\square I]}{\langle np_1 \rangle \circ_{1r} ((\langle x_1 \rangle \circ_{1l} np_2) \circ_2 v_2) \Rightarrow \square s}$$

We illustrate the semantic composition with (74). The derivational semantics is given in (75a). Applying the extended meaning of ‘ghoghho’ (rejoice) in (75c) to the representation of VP_2 and the two arguments of V_1 yields (75b).

(74) Òzó ghòghò ègiè khuòmwin.
 Ozo be-happy title be-sick
 ‘Ozo became sick after rejoicing over his title.’
 Stewart (2001:77)

(75) a. $((x_{v_1} x_{v_2}) x_{np_2}) x_{np_1}$.
 b. $\lambda e. \exists e_1. \exists e_2 [e = e_1 \wedge rejoice(e_1) \wedge be-sick(e_2) \wedge actor(e_1) = ozo \wedge theme(e_1) = \iota z. title(z) \wedge actor(e_1) = theme(e_2) \wedge e_1 \preceq e_2]$.
 c. $\lambda VP. \lambda y. \lambda x. \lambda e. \exists e_1. \exists e_2 [e = e_1 \wedge rejoice(e_1) \wedge VP(x)(e_2) \wedge actor(e_1) = x \wedge theme(e_1) = y \wedge actor(e_1) = first(e_2) \wedge e_1 \preceq e_2]$.

For reasons of symmetry to transitive verbs, an extended intransitive verb is assigned the type $\square(vp/i\ vp)$, i.e. the extension of the subcategorization frame is of type vp and the modal decoration is the same as that for the unextended verb.²⁰ With this assignment one gets the following derivation for a CC consisting of two intransitive verbs.

CC (two intransitive verbs):

$$\frac{\frac{\frac{np_1 \Rightarrow \square np}{\langle np_1 \rangle \Rightarrow np} [\square E] \quad \frac{\frac{v_1 \Rightarrow \square (vp/_2 vp)}{\langle v_1 \rangle \Rightarrow vp/_2 vp} [\square E] \quad \frac{v_2 \Rightarrow \square vp}{\langle v_2 \rangle \Rightarrow vp} [/_0 E]}{\langle v_1 \rangle \circ_2 \langle v_2 \rangle \Rightarrow vp} [\setminus_{1r} E]}{\langle np_1 \rangle \circ_{1r} (\langle v_1 \rangle \circ_2 \langle v_2 \rangle) \Rightarrow S} [K(\bullet_2)]}{\langle np_1 \rangle \circ_{1r} \langle v_1 \circ_2 v_2 \rangle \Rightarrow S} [K(\bullet_{1r})]}{\langle np_1 \circ_{1r} (v_1 \circ_2 v_2) \rangle \Rightarrow S} [\square I]}{np_1 \circ_{1r} (v_1 \circ_2 v_2) \Rightarrow \square S}$$

Note that the modal decoration of the extended verb with \square is exactly what is required. Since VP_1 consists only of V_1 , there being no

²⁰The situation is more complex since one has to take into account the fact that modification with a manner adverb before the second verb is inadmissible in an RSVC but not in a CSVC and a CC; see Section 4.8 below for details.

right-adjoined NP, $K(\bullet_2)$ removes the modal decoration of the linguistic resource corresponding to V_1 . If an extended intransitive verb were of type $\diamond\Box(vp/_2vp)$, this would lead to a sequent the antecedent term of which would not correspond to any substitution of lexical items (assuming the hypothesis $x_1 \Rightarrow \Box(vp/_2vp)$).

The derivation of RSVCs

4.7

The derivation of an RSVC has to take into account that in this type of SVC a manner adverb can occur only before the first but not before the second verb. Assuming that each position corresponds to a particular projection of the verb that is modified, manner adverbs require two such projections. For both the CSVC and the CC, there are subexpressions that are of type vp . The first corresponds to the VP built in terms of V_2 , which is the first argument of the (extended) verb V_1 . The second subexpression of type vp is that corresponding to the sequence $V_1 NP_2 V_2 (NP_3)$. Modification of this expression takes place in position 1.

If one takes a manner adverb in position 2 to modify VP_2 , i.e. the VP with head V_2 , the task consists in explaining why modification of this VP is possible in the context of an CSVC and a CC but not in the context of an RSVC. One strategy to explain this phenomenon is to use the unary connectives from the underlying logic. Recall that these connectives basically have two functions. They can either be used to license operations that are not available in the base logic or they can be used to restrict operations that are by default available in this logic. Theoretically, either of the two functions can be used to interpret the distribution of adverbs. In this article the second strategy will be adopted.

Manner adverbs are basically of type $vp/_a vp$ or $vp \setminus_a vp$.²¹ In order to block modification with an adverb, the second verb in an RSVC must be of a modally decorated type. Since the default type assigned to intransitive verbs is $\Box vp$, it has to be decorated differently. Suppose one makes the following assumptions in the context of an RSVC. The

²¹ \cdot_a is the adverbial adjunction mode that combines a verbal (phrasal) structure with an adverb.

head adjunction mode is \cdot_2 , i.e. the same mode that is used for a CC. The type of an intransitive second verb is $\square\square vp$ whereas that of extended intransitive verbs is $\square(vp/2\square\square vp)$. An extended transitive verb has type $\diamond\square(tv/2\square\square vp)$ and its unextended variants that occur as the second verb have type $\diamond\square(\square\square vp/1t np)$. Below the derivations for the three types of an RSVC are given. The derivation of an RSVC with two transitive verbs is displayed on page 387 and that with a transitive and an intransitive verb on page 388.

RSVC (two intransitive verbs):

$$\begin{array}{c}
 \frac{np_1 \Rightarrow \square np \quad [\square E]}{\langle np_1 \rangle \Rightarrow np} \quad \frac{v_1 \Rightarrow \square (vp/2 \square \square vp) \quad [\square E]}{\langle v_1 \rangle \Rightarrow vp/2 \square \square vp} \quad \frac{v_2 \Rightarrow \square \square vp \quad [\square E]}{\langle v_2 \rangle \Rightarrow \square \square vp} \\
 \hline
 \frac{\langle v_1 \rangle \circ_2 \langle v_2 \rangle \Rightarrow vp \quad [\cdot_2 E]}{\langle np_1 \rangle \circ_{1r} (\langle v_1 \rangle \circ_2 \langle v_2 \rangle) \Rightarrow S} \quad [K(\bullet_2)] \\
 \frac{\langle np_1 \rangle \circ_{1r} \langle v_1 \circ_2 v_2 \rangle \Rightarrow S}{\langle np_1 \circ_{1r} (v_1 \circ_2 v_2) \rangle \Rightarrow S} \quad [K(\bullet_{1r})] \\
 \frac{\langle np_1 \circ_{1r} (v_1 \circ_2 v_2) \rangle \Rightarrow S}{np_1 \circ_{1r} (v_1 \circ_2 v_2) \Rightarrow \square S} \quad [\square I]
 \end{array}$$

The case of two transitive verbs is illustrated with (76). The derivational semantics is given in (77a): the meaning representation of the extended V_1 ‘gbe’ (hit) in (77c) applied to VP_2 and the two arguments of V_1 yields (77b).

- (76) Òzó gbé èkhù làá òwá.
 Ozo hit door enter house
 ‘Ozo hit the door into the house.’
 Stewart (2001:145)

- (77) a. $((x_{v_1}(x_{v_2}x_{np_3})x_{np_2})x_{np_1})$.
 b. $\lambda e.\exists e_1.\exists e_2[e = e_1 \sqcup e_2 \wedge hit(e_1) \wedge enter(e_2) \wedge actor(e_1) = ozo \wedge theme(e_1) = \iota w.door(w) \wedge theme(e_1) = actor(e_2) \wedge theme(e_2) = \iota z.house(z) \wedge cause(e_1, e_2)]$.
 c. $\lambda VP_2.\lambda y.\lambda x.\lambda e.\exists e_1.\exists e_2[e = e_1 \sqcup e_2 \wedge hit(e_1) \wedge VP_2(y)(e_2) \wedge actor(e_1) = x \wedge theme(e_1) = y \wedge theme(e_1) = first(e_2) \wedge cause(e_1, e_2)]$.

The semantics for an RSVC with a transitive and an intransitive verb is illustrated with (78). The derivational semantics applied to the example is given in (79).

RSVC (two transitive verbs):

$$\begin{array}{c}
 \frac{[x_2 \Rightarrow \square(\square \square \text{vp} / \text{I} \text{np})]^2}{\langle x_2 \rangle \Rightarrow \square \square \text{vp} / \text{I} \text{np}} \frac{\text{np}_3 \Rightarrow \square \text{np}}{\langle \text{np}_3 \rangle \Rightarrow \text{np}} \frac{[\square \text{E}]}{[/_{\text{I} \text{E}}]} \\
 \frac{[x_1 \Rightarrow \square(\text{tv} / \text{I} \square \square \text{vp})]^1}{\langle x_1 \rangle \Rightarrow (\text{tv} / \text{I} \square \square \text{vp})} \frac{[\square \text{E}]}{\langle x_2 \rangle \circ_{\text{I} \text{I}} (\text{np}_3) \Rightarrow \square \square \text{vp}} \frac{\text{np}_2 \Rightarrow \square \text{np}}{\langle \text{np}_2 \rangle \Rightarrow \text{np}} \frac{[\square \text{E}]}{[/_{\text{I} \text{E}}]} \\
 \frac{[\square \text{E}]}{\langle x_1 \rangle \circ_2 (\langle x_2 \rangle \circ_{\text{I} \text{I}} (\text{np}_3)) \Rightarrow \text{tv}} \\
 \frac{[\square \text{E}]}{\langle x_1 \rangle \circ_2 (\langle x_2 \rangle \circ_{\text{I} \text{I}} (\text{np}_3)) \circ_{\text{I} \text{I}} (\text{np}_2) \Rightarrow \text{vp}} \frac{[\text{MP1}]}{\langle x_1 \rangle \circ_{\text{I} \text{I}} (\text{np}_2) \circ_2 (\langle x_2 \rangle \circ_{\text{I} \text{I}} (\text{np}_3)) \Rightarrow \text{vp}} \frac{[\text{K}^* 2(\bullet_{\text{I} \text{I}})]}{\langle x_1 \rangle \circ_{\text{I} \text{I}} \text{np}_2 \circ_2 (\langle x_2 \rangle \circ_{\text{I} \text{I}} \text{np}_3) \Rightarrow \text{vp}} \frac{[\text{K}(\bullet_{\bullet})]}{\langle x_1 \rangle \circ_{\text{I} \text{I}} \text{np}_2 \circ_2 (\langle x_2 \rangle \circ_{\text{I} \text{I}} \text{np}_3) \Rightarrow \text{vp}} \frac{[\wedge_{\text{I} \text{r}, \text{E}}]}{\langle \text{np}_1 \rangle \circ_{\text{I} \text{r}} (\langle x_1 \rangle \circ_{\text{I} \text{I}} \text{np}_2) \circ_2 (\langle x_2 \rangle \circ_{\text{I} \text{I}} \text{np}_3) \Rightarrow \text{s}} \frac{[\text{K}(\bullet_{\text{I} \text{r}})]}{\langle \text{np}_1 \circ_{\text{I} \text{r}} (\langle x_1 \rangle \circ_{\text{I} \text{I}} \text{np}_2) \circ_2 (\langle x_2 \rangle \circ_{\text{I} \text{I}} \text{np}_3) \rangle \Rightarrow \text{s}} \frac{[\square \text{I}]}{\text{np}_1 \circ_{\text{I} \text{r}} (\langle x_1 \rangle \circ_{\text{I} \text{I}} \text{np}_2) \circ_2 (\langle x_2 \rangle \circ_{\text{I} \text{I}} \text{np}_3) \Rightarrow \square \text{s}} \frac{v_1 \Rightarrow \diamond \square(\text{tv} / \text{I} \square \square \text{vp})}{\text{np}_1 \circ_{\text{I} \text{r}} ((v_1 \circ_{\text{I} \text{I}} \text{np}_2) \circ_2 (\langle x_2 \rangle \circ_{\text{I} \text{I}} \text{np}_3)) \Rightarrow \square \text{s}} \frac{[\diamond \text{E}]^1}{[\diamond \text{E}]^2} \\
 \frac{v_2 \Rightarrow \diamond \square(\square \square \text{vp} / \text{I} \text{np})}{\text{np}_1 \circ_{\text{I} \text{r}} ((v_1 \circ_{\text{I} \text{I}} \text{np}_2) \circ_2 (v_2 \circ_{\text{I} \text{I}} \text{np}_3)) \Rightarrow \square \text{s}} \frac{[\diamond \text{E}]^2}{[\diamond \text{E}]^2}
 \end{array}$$

RSVC (transitive and intransitive verb):

$$\begin{array}{c}
 \frac{[x_1 \Rightarrow \Box(tv/_2 \Box \Box vp)]^1}{\langle x_1 \rangle \Rightarrow (tv/_2 \Box \Box vp)} \quad \frac{[x_1] \quad v_2 \Rightarrow \Box \Box \Box vp}{\langle v_2 \rangle \Rightarrow \Box \Box \Box vp} \quad \frac{[\Box E] \quad [x_1] \quad np_2 \Rightarrow \Box np}{[/_2 E] \quad \langle np_2 \rangle \Rightarrow np} \quad \frac{[\Box E] \quad [/_1 E]}{[/_1 E]} \\
 \frac{\langle x_1 \rangle \circ_2 \langle v_2 \rangle \Rightarrow tv}{\langle x_1 \rangle \circ_2 \langle v_2 \rangle \circ_{1l} \langle np_2 \rangle \Rightarrow vp} \quad \frac{[MP1]}{(\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle) \circ_2 \langle v_2 \rangle \Rightarrow vp} \quad \frac{[K^*2(\bullet_{1l})]}{\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle \circ_2 \langle v_2 \rangle \Rightarrow vp} \quad \frac{[K(\bullet_2)]}{\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle \circ_2 \langle v_2 \rangle \Rightarrow vp} \quad \frac{[/_1 E]}{[/_1 E]} \\
 \frac{np_1 \Rightarrow \Box np}{\langle np_1 \rangle \Rightarrow np} \quad \frac{[\Box E]}{[\Box E]} \quad \frac{\langle np_1 \rangle \circ_{1r} ((\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle) \circ_2 v_2) \Rightarrow s}{\langle np_1 \rangle \circ_{1r} ((\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle) \circ_2 v_2) \Rightarrow s} \quad \frac{[K(\bullet_{1r})]}{[K(\bullet_{1r})]} \\
 \frac{np_1 \circ_{1r} ((\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle) \circ_2 v_2) \Rightarrow \Box s}{\langle np_1 \rangle \circ_{1r} ((\langle x_1 \rangle \circ_{1l} \langle np_2 \rangle) \circ_2 v_2) \Rightarrow \Box s} \quad \frac{[[\Box]]}{[[\Box]]} \\
 \frac{v_1 \Rightarrow \Diamond \Box (tv/_2 \Box \Box vp)}{np_1 \circ_{1r} (v_1 \circ_{1l} \langle np_2 \rangle \circ_2 v_2) \Rightarrow \Box s} \quad \frac{[\Diamond E]^1}{[\Diamond E]^1}
 \end{array}$$

- (78) Òzó kòkó Àdésúwà m̀̀s̀̀é.
 Ozo raise Adesuwa be-beautiful
 ‘Ozo raised Adesuwa to be beautiful.’
 Stewart (2001:12)
- (79) a. $((x_{v_1} x_{v_2}) x_{np_2}) x_{np_1}$.
 b. $\lambda e. \exists e_1. \exists e_2 [e = e_1 \sqcup e_2 \wedge \text{raise}(e_1) \wedge \text{be_beautiful}(e_2) \wedge \text{actor}(e_1) = \text{ozo} \wedge \text{theme}(e_1) = \text{adusewa} \wedge \text{theme}(e_1) = \text{theme}(e_2) \wedge \text{cause}(e_1, e_2)]$
 c. $\lambda VP_2. \lambda y. \lambda x. \lambda e. \exists e_1. \exists e_2 [e = e_1 \sqcup e_2 \wedge \text{raise}(e_1) \wedge VP_2(y)(e_2) \wedge \text{actor}(e_1) = x \wedge \text{theme}(e_1) = y \wedge \text{theme}(e_1) = \text{first}(e_2) \wedge \text{cause}(e_1, e_2)]$

For an RSVC with two intransitive verbs, we consider (80).

- (80) Òzó dé wú.
 Ozo fall die
 ‘Ozo fell to death.’
 Stewart (2001:15)
- (81) a. $((x_{v_1} x_{v_2}) x_{np})$.
 b. $\lambda e. \exists e_1. \exists e_2 [e = e_1 \sqcup e_2 \wedge \text{fall}(e_1) \wedge \text{die}(e_2) \wedge \text{actor}(e_1) = \text{ozo} \wedge \text{actor}(e_1) = \text{theme}(e_2) \wedge \text{cause}(e_1, e_2)]$.
 c. $\lambda VP_2. \lambda x. \lambda e. \exists e_1. \exists e_2 [e = e_1 \sqcup e_2 \wedge \text{fall}(e_1) \wedge VP_2(x)(e_2) \wedge \text{actor}(e_1) = x \wedge \text{actor}(e_1) = \text{first}(e_2) \wedge \text{cause}(e_1, e_2)]$.

In contrast to a CSVC, the manner adverb ‘giegie’ cannot occur in position 2 of an RSVC. In the text this inadmissibility has been explained by a modal decoration at the syntactic level. One may argue that there is an alternative, semantic explanation. The inadmissibility of this type of adverb in position 2 results if one assumes that the VP headed by V_2 is not a constituent of the sentence. One way of achieving this is to assume that in an RSVC the complex predicate is not an extended verb that has an additional VP argument but a basic complex predicate. For example, the meaning of ‘de’ (fall) when used as first verb in an RSVC would be (82).

- (82) $\lambda y. \lambda x. \lambda e. \exists e_1. \exists e_2 [e = e_1 \sqcup e_2 \wedge \text{fall}(e_1) \wedge \text{die}(e_2) \wedge \text{actor}(e_1) = \text{first}(e_2) \wedge \text{theme}(e_1) = \text{second}(e_2) \wedge \text{cause}(e_1, e_2)]$.

Generalizing this argument, one may say that this strategy applies whenever all arguments of the second verb are shared with an argument of the first verb. From this perspective it also applies to a CSVC with two transitive verbs. However, this strategy faces the following two problems. First, in a CSVC with two transitive verbs a manner adverb can occur in position 2. This problem could be solved by assuming that 'giegie' can itself infix into a complex predicate. This means however that 'giegie' needs to be assigned an additional syntactic type and that an additional mechanism is necessary to explain why this infixation is blocked for an RSVC. The second problem is that this strategy cannot be applied if not all arguments of the second verb are shared with one argument of the first verb. This means that it cannot be applied to CSVCs with two ditransitive verbs (indirect objects must be different) and in RSVCs with two transitive verbs (direct objects need not be shared). Hence, this strategy fails to apply even to one subtype of an SVC without exception.

4.8

The derivation of CSVCs with ditransitive verbs

Similarly to a CSVC with two transitive verbs, in a CSVC with a ditransitive verb the subjects and direct objects are identified and the direct object of the second verb cannot be overtly realized. By contrast, the indirect object of the ditransitive verb is not identified with any object of the other verb. In particular, in the case of a CSVC with two ditransitive verbs, the indirect objects are not identified.

If, for a ditransitive verb, one assumes the order of arguments that are looked for to the right to be IO – DO, a ditransitive verb poses no problems at the level of word order since the objects are concatenated in the correct order: V NP_{IO} NP_{DO}. However, if the order is DO – IO, as this is assumed for instance in Lexical Decomposition Grammar (Gamerschlag 2005), one gets V NP_{DO} NP_{IO}. One strategy that has been applied to achieve the correct word order is the use of so-called discontinuity operators (see e.g. Morrill 1994, 1995). The functors built from the directional slashes adjoin either to the left or to the right of their arguments to form a continuous string. For functors built from a discontinuity operator, functor and argument are composed in a different way. The first sort of such operators are wrapping

and infixing operators. A functor $B\uparrow A$ wraps around an argument of type A to form a B . By contrast, a functor $B\downarrow A$ infixes itself in an A to form a B . In order to wrap around an A the functor expression must consist of two parts. For example, if these parts are s and s' , wrapping yields $s + s'' + s'$, for s'' being an expression of type A . The second sort of discontinuity operators are used to construe such ‘splitting’ or pair expressions. An expression of type $B<A$ takes an expression of type A to form a pair expression with the functor expression as first and the argument expression as second element: Using $<$ and \uparrow , a ditransitive verb can be assigned the type $(vp \uparrow np) < np$. Given an appropriate permutation rule, $vp/_l np_2 / np_1$ is derivable from $(vp \uparrow np_1) < np_2$.

In a multimodal variant of NL(\diamond) this strategy can be simulated in the following way. A wrapping or infixing operation is modelled by a permutation rule. The discontinuity operators can be represented by particular modes of composition. Moortgat and Oerhle (1993) distinguish four types of head wrapping modes: \cdot_{ij} with $i = 1l$ or $i = 1r$ and $j = h$ or $j = d$. The first index indicates the infix and the second index indicates whether the infix is the head (h) or the dependent (d) of the combination. The mixed permutation rule MP2 says that a left dependent infix (B) can be infixes in a \circ_{1l} structure.

$$(83) \quad \text{MP2: } (A \bullet_{rd} B) \bullet_{1l} C \rightarrow (A \bullet_{1l} C) \bullet_{rd} B$$

The relationship between \cdot_{1l} and \cdot_{1r} on the one hand and the head wrapping modes \cdot_{ij} is captured by rules such as that in (84).

$$(84) \quad \text{K}(l/rd): A \bullet_{1l} B \rightarrow A \bullet_{rd} B$$

Adopting this strategy, a ditransitive verb is assigned the types in (85).

$$(85) \quad \diamond\Box (vp/_rd np/_1l np) \text{ (unextended); } \diamond\Box (vp/_rd np/_1l np/_0 vp) \text{ (extended)}$$

In order to derive a simple sentence with a ditransitive verb needed are the two structural rules in (86).

$$(86) \quad \begin{array}{l} \text{a. } K^*(\bullet_{1l}): \diamond((\diamond A \bullet_{rd} B) \bullet_{1l} C) \rightarrow \diamond(\diamond A \bullet_{rd} B) \bullet_{1l} \diamond C \\ \text{b. } K^2(\bullet_{rd}): \diamond(\diamond A \bullet_{rd} B) \rightarrow \diamond A \bullet_{rd} \diamond B \end{array}$$

The rule $K^*(\bullet_{1l})$ allows for the percolation of the modal decorations of both components of a \circ_{1l} -structure if the left component is

a \circ_{rd} -structure, i.e. a structure which composes a (lexical) verbal element with an NP. Thus, this rule is applicable only in the context of ditransitive verbs. The rule $K^*(\bullet_{rd})$ is similar to the rule $K^*(\bullet_{1l})$. It allows for the percolation of the modal decoration of the right component of a \circ_{rd} -structure, provided its left component is modally decorated, too.

Derivation of the VP in a simple sentence with a ditransitive verb:

$$\begin{array}{c}
 \frac{x \Rightarrow \Box(vp/_{rd}np/_{1l}np)}{\langle x \rangle \Rightarrow vp/_{rd}np/_{1l}np} \quad \frac{np_3 \Rightarrow \Box np}{\langle np_3 \rangle \Rightarrow np} \quad \frac{np_2 \Rightarrow \Box np}{\langle np_2 \rangle \Rightarrow np} \\
 \frac{\frac{\frac{\frac{\frac{\langle x \rangle \circ_{1l} \langle np_3 \rangle \Rightarrow vp/_{rd}np}{\langle x \rangle \circ_{1l} \langle np_3 \rangle \circ_{rd} \langle np_2 \rangle \Rightarrow vp} \quad [MP2]}{\langle x \rangle \circ_{rd} \langle np_2 \rangle \circ_{1l} \langle np_3 \rangle \Rightarrow vp} \quad [K^*(\bullet_{rd})]}{\langle x \rangle \circ_{rd} np_2 \circ_{1l} \langle np_3 \rangle \Rightarrow vp} \quad [K^*(\bullet_{1l})]}{\langle \langle x \rangle \circ_{rd} np_2 \rangle \circ_{1l} np_3 \Rightarrow vp} \quad [K(l/rd)]]}{\langle \langle x \rangle \circ_{1l} np_2 \rangle \circ_{1l} np_3 \Rightarrow vp} \quad [MP1]
 \end{array}$$

Not applying MP2 has the same effect as in the case of MP1. If in line 6 $K^*1(\bullet_{1l})$ instead of $K^*(\bullet_{1l})$ is used, the structural operator of $\langle \langle x \rangle \circ_{rd} np_2 \rangle$ is not percolated. Since the semantics adds nothing new, it is skipped.

For the derivation of a CSVC with a ditransitive first and a transitive second verb, the mixed permutation rule MP3 is needed.

$$(87) \quad MP3: (A \bullet_{rd} C) \bullet_0 B \rightarrow (A \bullet_0 B) \bullet_{rd} C$$

Below the relevant steps of the derivation of the VP are given.

$$\begin{array}{c}
 \frac{((\langle x_1 \rangle \circ_0 (\langle x_2 \rangle \circ_{1l} \langle np_3 \rangle)) \circ_{1l} \langle np_3 \rangle) \circ_{rd} \langle np_2 \rangle \Rightarrow vp}{((\langle x_1 \rangle \circ_{1l} \langle np_3 \rangle) \circ_0 (\langle x_2 \rangle \circ_{1l} \langle np_3 \rangle)) \circ_{rd} \langle np_2 \rangle \Rightarrow vp} \quad [MP1] \\
 \frac{((\langle x_1 \rangle \circ_{1l} \langle np_3 \rangle) \circ_{rd} \langle np_2 \rangle) \circ_0 (\langle x_2 \rangle \circ_{1l} \langle np_3 \rangle) \Rightarrow vp}{((\langle x_1 \rangle \circ_{rd} \langle np_2 \rangle) \circ_{1l} \langle np_3 \rangle) \circ_0 (\langle x_2 \rangle \circ_{1l} \langle np_3 \rangle) \Rightarrow vp} \quad [MP2] \\
 \frac{((\langle x_1 \rangle \circ_{rd} \langle np_2 \rangle) \circ_0 \langle x_2 \rangle) \circ_{1l} \langle np_3 \rangle \Rightarrow vp}{((\langle x_1 \rangle \circ_{rd} \langle np_2 \rangle) \circ_{1l} \langle np_3 \rangle) \circ_0 \langle x_2 \rangle \Rightarrow vp} \quad [MC] \\
 \frac{((\langle x_1 \rangle \circ_{rd} \langle np_2 \rangle) \circ_{1l} \langle np_3 \rangle) \circ_0 \langle x_2 \rangle \Rightarrow vp}{((\langle x_1 \rangle \circ_{rd} np_2) \circ_{1l} \langle np_3 \rangle) \circ_0 \langle x_2 \rangle \Rightarrow vp} \quad [MP1] \\
 \frac{((\langle x_1 \rangle \circ_{rd} np_2) \circ_{1l} \langle np_3 \rangle) \circ_0 \langle x_2 \rangle \Rightarrow vp}{\langle \langle x_1 \rangle \circ_{rd} np_2 \rangle \circ_{1l} np_3 \circ_0 \langle x_2 \rangle \Rightarrow vp} \quad [K^*2(\bullet_{rd})] \\
 \frac{\langle \langle x_1 \rangle \circ_{rd} np_2 \rangle \circ_{1l} np_3 \circ_0 \langle x_2 \rangle \Rightarrow vp}{\langle \langle \langle x_1 \rangle \circ_{rd} np_2 \rangle \circ_{1l} np_3 \rangle \circ_0 \langle x_2 \rangle \Rightarrow vp} \quad [K^*(\bullet_{1l})] \\
 \frac{\langle \langle \langle x_1 \rangle \circ_{rd} np_2 \rangle \circ_{1l} np_3 \rangle \circ_0 \langle x_2 \rangle \Rightarrow vp}{\langle \langle \langle x_1 \rangle \circ_{1l} np_2 \rangle \circ_{1l} np_3 \rangle \circ_0 \langle x_2 \rangle \Rightarrow vp} \quad [K1(\bullet_0)] \\
 \frac{\langle \langle \langle x_1 \rangle \circ_{1l} np_2 \rangle \circ_{1l} np_3 \rangle \circ_0 \langle x_2 \rangle \Rightarrow vp}{\langle \langle \langle x_1 \rangle \circ_{1l} np_2 \rangle \circ_{1l} np_3 \rangle \circ_0 \langle x_2 \rangle \Rightarrow vp} \quad [K(l/rd)]
 \end{array}$$

The by now familiar arguments apply if particular rules are not used or if the order is reversed. For example, if MC is not applied, one

only gets a structure of the form $\langle \Gamma \rangle \circ_0 (\langle x_2 \rangle \circ_{1l} \langle np_3 \rangle)$. The structural operator from np_3 must be percolated. Yet this is not possible because $K1(\bullet_0)$ only percolates the structural operator of the left component. If MP1 is not applied in line 5, one gets the following continuation.

$$\frac{((\langle x_1 \rangle \circ_{rd} \langle np_2 \rangle) \circ_0 \langle x_2 \rangle) \circ_{1l} \langle np_3 \rangle \Rightarrow vp}{((\langle x_1 \rangle \circ_{rd} np_2) \circ_0 \langle x_2 \rangle) \circ_{1l} \langle np_3 \rangle \Rightarrow vp} [K^*2(\bullet_{rd})]$$

$$\frac{((\langle x_1 \rangle \circ_{rd} np_2) \circ_0 \langle x_2 \rangle) \circ_{1l} \langle np_3 \rangle \Rightarrow vp}{\langle (\langle x_1 \rangle \circ_{rd} np_2) \circ_0 \langle x_2 \rangle \rangle \circ_{1l} \langle np_3 \rangle \Rightarrow vp} [K1(\bullet_0)]$$

Now only rule $K^*2(\bullet_{1l})$ can be used, which does not percolate the structural operator of the left component. Yet, this operator has to be percolated since it originates from np_2 . An analogous argument applies if in line 7 instead of $K^*(\bullet_{1l})$ $K^*2(\bullet_{1l})$ is used.

Skipping the application of the structural rule for the subject, we will give the semantic derivation for (88).

- (88) Úyi hàè Ìsòkèṅ íghó dó-rhié
 Uyi pay Isoken money steal
 ‘Uyi paid Isoken the money and stole it.’
 Stewart (2001:137)

- (89) a. $((((x_{v_1}(x_{v_2}x_{np_3}))x_{np_3})x_{np_2})x_{np_1})$
 b. $\lambda e.\exists e_1.\exists e_2[e = e_1 \sqcup e_2 \wedge pay(e_1) \wedge steal(e_2) \wedge actor(e_1) = uyi \wedge theme(e_1) = \iota w.money(w) \wedge goal(e_1) = isoken \wedge actor(e_1) = actor(e_2) \wedge theme(e_1) = theme(e_2) \wedge e_1 \preceq e_2 \wedge \Box_{uyi}(occur(e_1) \rightarrow occur(e_2))]$
 c. $\lambda VP_2.\lambda z.\lambda y.\lambda x.\lambda e.\exists e_1.\exists e_2[e = e_1 \sqcup e_2 \wedge pay(e_1) \wedge VP_2(x)(e_2) \wedge actor(e_1) = x \wedge theme(e_1) = z \wedge goal(e_1) = y \wedge actor(e_1) = first(e_2) \wedge theme(e_1) = second(e_2) \wedge e_1 \preceq e_2 \wedge \Box_x(occur(e_1) \rightarrow occur(e_2))]$

For a CSVC with a transitive first and a ditransitive second verb, the relevant steps of the derivation of the VP are shown below.

$$\frac{\langle x_1 \rangle \circ_0 ((\langle x \rangle \circ_{rd} \langle np_2 \rangle) \circ_{1l} \langle np_3 \rangle) \circ_{1l} \langle np_3 \rangle \Rightarrow vp}{\langle x_1 \rangle \circ_{1l} \langle np_3 \rangle \circ_0 ((\langle x \rangle \circ_{rd} \langle np_2 \rangle) \circ_{1l} \langle np_3 \rangle) \Rightarrow vp} [MP1]$$

$$\frac{\langle x_1 \rangle \circ_0 ((\langle x \rangle \circ_{rd} \langle np_2 \rangle) \circ_{1l} \langle np_3 \rangle) \Rightarrow vp}{\langle x_1 \rangle \circ_{1l} \langle np_3 \rangle \circ_0 ((\langle x \rangle \circ_{rd} \langle np_2 \rangle) \Rightarrow vp} [MC]$$

$$\frac{\langle x_1 \rangle \circ_{1l} \langle np_3 \rangle \circ_0 ((\langle x \rangle \circ_{rd} \langle np_2 \rangle) \Rightarrow vp}{\langle x_1 \rangle \circ_{1l} np_3 \circ_0 \langle x \rangle \circ_{rd} \langle np_2 \rangle \Rightarrow vp} [MP1]$$

$$\frac{\langle x_1 \rangle \circ_{1l} np_3 \circ_0 \langle x \rangle \circ_{rd} \langle np_2 \rangle \Rightarrow vp}{\langle x_1 \rangle \circ_{1l} np_3 \circ_0 \langle x \rangle \circ_{rd} np_2 \Rightarrow vp} [K^*2(\bullet_{1l})]$$

$$\frac{\langle x_1 \rangle \circ_{1l} np_3 \circ_0 \langle x \rangle \circ_{rd} np_2 \Rightarrow vp}{\langle x_1 \rangle \circ_{1l} np_3 \circ_0 \langle x \rangle \circ_{rd} np_2 \Rightarrow vp} [K^*2(\bullet_{rd})]$$

Now a problem arises because $K1(\bullet_0)$ only percolates the structural operator of the left component and leaves the right component unchanged. Yet, in this particular case the structural operator of the left component has to be percolated, too. Noticing that the right structure is composed by \circ_{rd} , this problem can be overcome by adding the rule $K^*(\bullet_0)$.

$$(90) \quad K^*(\bullet_0): \diamond(A \circ_0 (\diamond B \circ_{rd} C)) \rightarrow \diamond A \circ_0 \diamond(\diamond B \circ_{rd} C)$$

$K^*(\bullet_0)$ is applicable only in the context of a verbal cluster with a ditransitive verb to which MC has been applied. Using this rule, one gets line 7.

$$7. \quad \langle\langle x_1 \rangle \circ_{1l} np_3 \rangle \circ_0 (\langle x \rangle \circ_{rd} np_2) \Rightarrow vp$$

Applying $K1(\bullet_0)$ in line 6 does not percolate the structural operator originating from np_3 . If MPI is not used in line 3, the structural operator of this resource is likewise not percolated. If MC is not applied in line 2, it is possible to derive the sequent in (91) by applying $K^*2(\bullet_{rd})$ and $K^*2(\bullet_{1l})$ to the left component of this line.

$$(91) \quad \langle\langle x_1 \rangle \circ_{1l} \langle np_3 \rangle \rangle \circ_0 \langle\langle x \rangle \circ_{rd} np_2 \rangle \circ_{1l} np_3 \Rightarrow vp$$

$K^*(\bullet_0)$ can be applied to this sequent. Yet since the structural operator of the left component of $\langle\langle x \rangle \circ_{rd} np_2 \rangle \circ_{1l} np_3$ is not percolated, the sequent is linguistically ill-formed. If instead of $K^*2(\bullet_{1l})$ $K^*(\bullet_{1l})$ is used, one gets the sequent in (92).

$$(92) \quad \langle\langle x_1 \rangle \circ_{1l} \langle np_3 \rangle \rangle \circ_0 \langle\langle x \rangle \circ_{rd} np_2 \rangle \circ_{1l} np_3 \Rightarrow vp$$

Though this removes the structural operator of the left component of $\langle\langle x \rangle \circ_{rd} np_2 \rangle \circ_{1l} np_3$, now rule $K^*(\bullet_0)$ cannot be applied because it requires this left component to be modally decorated. Application of rule $K1(\bullet_0)$ only percolates the structural operator of the left but not that of the right component. Yet, both operators must be percolated to the dominating \circ_0 -structure.

4.9

A sketch of an analysis of manner adverbs

Due to space restrictions we cannot give a detailed analysis of manner adverbs. Manner adverbs are basically of syntactic type $vp/_a vp$ or $vp \setminus_a vp$ with \cdot_a the adverbial adjunction mode that combines a verbal

(phrasal) structure with an adverb. Hence, there is nothing new compared to standard analyzes of adverbs in other languages. In an SVC or a CC there are two VPs. One is projected by V_2 and the other is projected by the extended verb V_1 . In position 2 the adverb modifies the VP projected by V_2 whereas in position 1 it is the VP projected by V_1 that gets modified. Since V_2 is interpreted relative to e_2 , it is this event that is ascribed the property expressed by the adverb. By contrast, if the VP projected by V_1 is modified, the property is ascribed to the event denoted by the complex predicate. In an SVC this is the sum event $e = e_1 \sqcup e_2$ whereas in a CC it is e_1 .

COMPARISON TO OTHER APPROACHES

5

A comparison to Baker and Stewart 1999 and 2001

5.1

The analysis in Baker and Stewart (1999) is based on two assumptions. Following Hale and Keyser (1993), they assume that (canonical)²² transitive verbs semantically decompose into a causal/process and a transition/result component. This bipartition at the semantic level is reflected in the syntax by distinguishing between a v and a V element, with the former corresponding to the causal/process and the latter corresponding to the transition/result component. In addition to this distinction, it is assumed that agentive subjects are generated in the specifier position of a Voice Phrase (Kratzer 1996). The dominance relation is $\text{Voice} > v > V$. The three multiverb sequences are then distinguished in terms of the types of nodes that are independently projected by the two component verbs.

- (93) a. RSVC: there are no independent projections common to both verbs. Rather, since V_1 is a (canonical) transitive verb, it has both a v and a V component. In an RSVC, this VP does not immediately dominate V but V' , which,

²² An example for non-canonical transitive verbs given by Baker and Stewart (1999:18) are stative verbs, which are not admissible as the first verb in an RSVC and a CSVC.

in turn, immediately dominates V_1 and V_2 (Baker and Stewart 1999:18). Consequently, there is only one VP, one vP and one VoiceP.

- b. CSVC: each verb projects its own VP and vP. Since vP is the highest node independently projected by a component verb, the two verbs are merged at the level of vP. As a result, one has two VPs but three vPs: vP_1 , vP_2 and $vP_{1/2}$, which immediately dominates both vP_1 and vP_2 .
- c. CC: each verb projects its own VP, vP and VoiceP. Consequently, there are two VPs and two vPs. Since VoiceP is the maximal node independently projected by a component verb, the maximal projections of the verbs are merged at the level of VoiceP so that there are three nodes of this type: $VoiceP_1$, $VoiceP_2$ and $VoiceP_{1/2}$, the latter immediately dominating both $VoiceP_1$ and $VoiceP_2$.

Since both in a CSVC and a CC the two component verbs are treated on a par in the sense that each verb projects the same types of nodes, it follows that there should be no asymmetries in the interpretation of adverbs. Yet this is not the case. Manner adverbs like 'giegie' (quickly) behave asymmetrically in a CSVC. Before the first verb, it is the joint action expressed by both verbs that is required to have the property expressed by the adverb whereas an adverb of this type between NP_2 and the second verb imposes this requirement only on the action expressed by the second verb. According to Baker and Stewart (1999, 2001), adverbs like 'giegie' can be attached either to VoiceP or to vP, but not to VP. The authors account for the interpretation of those adverbs before the second verb by attaching it to $vP_{1/2}$, i.e. the vP node at which the two projections are merged in a CSVC. Consequently, both events (or their join) must be semantically accessible at this node. By contrast, attaching an adverb of this type to vP_2 accounts for the interpretation before the second verb according to which only the action expressed by V_2 is required to have the property. The problem now is that, by symmetry, an adverb of this type should also be attachable to vP_1 , yielding the interpretation that it is the action expressed by V_1 which has the corresponding property. Yet, an adverb like 'giegie' does not have such an interpretation. An analogous prob-

lem arises for adverbially modified CCs. A similar criticism applies to Stewart (2001).

Thus, in an analysis which treats both verbs on a par, an adverb that attaches to XP such that there can be up to three nodes of this type in an SVC or a CC should (i) induce three different interpretations and (ii) have the same interpretations relative to V_1 and V_2 . Both predictions are not borne out by manner adverbs like ‘giegie’. By contrast, in our analysis these adverbs always modify expressions of type *vp*.²³ Since the two component verbs are treated asymmetrically, only two subexpressions of type *vp* are generated. One is headed by the unextended second verb whereas the second is projected by the extended first verb.

The approach of Ogie 2010

5.2

In contrast to Baker and Stewart, Ogie (2010) does not analyze CSVCs in terms of *pro* in the object position of V_2 . Working in the HPSG framework and following Hellan *et al.* (2003), she bases her analysis on a distinction between different types of argument sharing patterns. The first pattern is token sharing by grammatical functions. In this pattern the verbs $V_1 \dots V_n$ share an NP token that is syntactically realized as an argument of V_1 . As an effect, there is one token NP bearing a particular grammatical function to the verbs in the series. This pattern is used for subjects and objects in a CSVC. At the formal level, this pattern is represented as identity between the values of the QVAL attribute of the head-daughter and the non-head-daughter with the token being instantiated on the VAL list of the head-daughter. For an RSVC, token sharing by grammatical function is not possible because in this pattern two argument positions share all (grammatical) properties. This constraint on token sharing does not hold in an RSVC simply because the argument is assigned the grammatical function of direct object relative to V_1 and subject relative to V_2 . Hence, the argument sharing pattern must be different. For an RSVC, the pattern is switch sharing.

²³Note that we follow the conventions of Type Logical Grammar in using lower case letters for maximal projections of lexical heads. In this sense ‘*vp*’ is headed by a verb and must not be confused with ‘*vp*’ projected by a head such as ‘*cause*’ in present day generative syntax.

In this pattern, the NP which bears the grammatical function of direct object to V_1 and which is overtly realized in its canonical position also bears the subject function to V_2 . Formally, this is represented by identifying the referential index of the non-head-daughter SUBJECT value with the value of the head-daughter's DOBJ's value. For the subjects in a CC, the argument sharing pattern is that of covert reference sharing. In this pattern, the NP which bears the grammatical function of subject to V_1 shares its referential index with the unsaturated subject argument of V_2 . A subject is unsaturated if it is not realized on the valence list of the verb to which it bears this grammatical function. At the formal level the value of the SUBJECT attribute is identified with the XARG value for the non-head-daughter. The non-head-daughter's XARG value is in turn identified with its SUBJECT's INDEX value by identifying the referential index.

Ogie uses the distribution of the 'tobore' anaphora as empirical evidence for her assigning of argument sharing patterns. This anaphora is used for emphasis and its basic use is as a subject oriented adverb. Importantly, it cannot occur in object position. For CSVCs, CCs and RSVCs, one gets the following pattern (Ogie 2010:295).²⁴

- (94) a. *Òzó_k lé èvbàrè tòbòrè_k ré.
 Ozo cook food by.himself eat
 intended: 'Ozo cooked food and ate it by himself.' CSVC
- b. Òzó_k dẹ ízẹ tòbòrè_k rí ọré.
 Ozo buy rice by.himself ate it
 Ozo bought rice and ate it by himself.' CC
- c. *Òzó_k kòkó Àdésúwà tòbòrè_k mọsẹ.
 Ozo raise Adesuwa by.himself be.beautiful
 intended: 'Ozo raised Adesuwa by himself to be beautiful.' RSVC

These examples show that 'tobore' is admissible before V_2 only in the CC construction. Having three argument sharing patterns in place, Ogie analyzes the distribution of the anaphora 'tobore' as follows, (Ogie 2010:302). Clauses in which this anaphora is not licensed before

²⁴For the sake of simplicity, we have reduced the more detailed glosses by Ogie in (94) and (95).

V_2 are analyzed as having one token NP bearing the subject grammatical function of the verbs in the construction. By contrast, clauses in which ‘tobore’ can occur before V_2 are analyzed as sharing referents between the subject arguments of the verbs in the series and $V_2 \dots V_n$ have unsaturated subjects. When taken at face value this explanation only accounts for the cases of CSVCs and CCs but not for the case of an RSVC. Ogie is aware of this and adds that a second type of clause, prohibiting the anaphora before V_2 , is characterized by the switch argument pattern.

However, this move is not convincing because it brings about the question as to what is the property common to the token sharing pattern and the switch sharing pattern that sets them aside from the overt reference sharing pattern underlying a CC. This property cannot be token sharing because this requires identity of grammatical function, a requirement that is not met in an RSVC where the direct object of V_1 is related to the subject of V_2 . Recall that in an RSVC the switch sharing is realized by identity of the referential index between the direct object of V_1 and the subject of V_2 . One possibility is to assume that token sharing by grammatical function implies identity of their corresponding referential indices. As an effect, this latter property would be common to the two argument sharing patterns characterizing the two types of SVCs. The problem with this explanation is that identity of the referential indices is also used for the pattern of overt reference sharing. Hence, one has to conclude that identity of referential indices cannot be the common property of the argument patterns underlying SVCs that explains the distribution of ‘tobore’.

Ogie defines the relation between the events denoted by SVCs and CCs in terms of the temporal relation between them. Two relations are distinguished. Disjointness of two events requires that the first event (completely) precedes the second. Two events are partially ordered if they are disjoint and if, in addition, the second event occurs immediately after the first (e_1 meets e_2). Whereas disjointness characterizes the relation between the events both in CSVCs and in CCs, events denoted by RSVCs are related by the partial order relation. From these definitions it follows that Ogie does not define the difference between SVCs and CCs at the level of single vs. non-single (join) of events. This has the effect that there is no difference between CSVCs and CCs at the level of events because the relation between the events is reduced

to the temporal relation between them. However, this does not capture the constraint on the events denoted by a CSVC that the actor carries out the first event with the intention to carry out the second event afterwards. Furthermore, it is not captured that manner adverbs in position 1, i.e. before V_1 are interpreted as determining a property of the sum of the events and not only of the event contributed by the interpretation of the first verb. By contrast, in our approach SVCs are semantically characterized by the fact that the complex predicate is interpreted relative to the sum of the events. As a result, manner adverbs in position 1 are interpreted with respect to this sum, in accordance with the data.

A third criticism has to do with the question of whether Ogie's analysis of the distribution of 'tobore' generalizes to other kinds of expressions which show a particular distributional pattern in SVCs and CCs, like manner adverbs, for example. Her analysis of 'tobore' does not directly generalize to this class of adverbs since they are not syntactically related to an NP but to a verb or the VP headed by it. In particular, the adverb applies to VP_2 before the modified VP combines with the extended verb related to V_1 both in a CSVC and a CC. It does, therefore, play no role whether the subject of V_2 is 'unsaturated' or whether it is token-identical to the subject of V_1 . Hence, Ogie needs a different mechanism to explain the distribution of manner adverbs.

A final question is the following: what is the relation between the templates for SVCs and CCs on the one hand and that for verbs in simple sentences on the other? It seems that different entries are required depending on whether the verb occurs as the second verb in an SVC or in a CC. For example, in a CC the subject of V_2 is unsaturated whereas in a CSVC this is not the case. In our approach verbs that can occur as the first verb in an SVC or CC have different types.

Let us compare Ogie's approach with ours. Ogie develops her analysis at the level of argument sharing patterns. In contrast to this approach, argument sharing patterns are not used to explain differences between RSVCs, CSVCs and CCs. Rather these differences are explained as differences at the semantic level and, hence, at the level of event structure. But even at the level of argument sharing patterns the analyses differ. In our approach, there is no difference between token and reference sharing. For example, if two arguments are shared, this means that they are 'token-identical' in the sense that there is a

single referent that bears the thematic relation(s) to the two events.

We will close by discussing an example involving quantification. In Ogie's approach, one effect of token sharing by grammatical function is that it ensures that all properties of the NP, including scope resolution with V_2 in an adjunction relation to V_1 , are shared. This becomes relevant for the interpretation of the two examples below (Ogie 2010:416).

- (95) a. Òzó dẹ èbé khéré tié.
Ozo buy book few read
'Ozo bought a few books and read them (all).' CSVC
- b. Òzó sùá èrhán khéré dè-lé.
Ozo push tree few fall
'Ozo pushed a few trees down.' RSVC

Baker and Stewart (2002) observed that (95a) has an E-type reading. It is true only if Ozo bought a few books in total and read them all. By contrast, (95b) is true in a situation in which Ozo pushed many trees but only a few fell as an effect of the pushing. Ogie (2010:417) argues that the interpretation of (95a) follows from the fact that due to token sharing of the objects the quantifier has scope over both verbs since all properties are shared. By contrast, in the RSVC the switch sharing pattern applies. This pattern involves different grammatical functions so that the scopal properties are not shared. As an effect the quantifier has scope only over V_2 .

Though we cannot give an account of quantification in this article, mainly due to the fact that this requires an extension of compositional semantics and event semantics along the lines proposed in Champollion (2015) and Bott and Sternefeld (2017), we will sketch how the above data can be analyzed in our approach. So far we assumed that there is a single event that is targeted, via λ -abstraction, in a complex predicate. For SVCs, this is the join $e = e_1 \sqcup e_2$ of the events in the action sequence whereas it is only the first event e_1 in this sequence in a CC. Data like (95) show that the actual situation is more complex. There need not be a single event that is targeted by operators that take the complex predicate as argument. Rather, which event is targeted depends on the operator. One way to account for this dependency on the operator is to interpret complex predicates relative to sets of events. As a result, the operator can 'select' one event in this set. We

assume the following selection criteria. For manner adverbs: the maximal event relative to \sqsubseteq in this set is selected, and for quantifiers like ‘khere’: the first event in the action sequence that is minimal relative to \sqsubseteq is selected. Using these two criteria, one can set $E = \{e, e_1, e_2\}$ as the most general solution, i.e. each complex predicate makes both the single events and their join accessible for operations. However, given the fact that e_1 is always targeted in a CC and e in an RSVC, for both operations considered here it is possible to restrict the choices in the following way. For a CSVC: $E = \{e, e_1\}$, for an RSVC: $E = \{e\}$ and for a CC: $E = \{e_1\}$.

The distribution of ‘tobore’ can equally be explained by a selection criterion. It selects the maximal event in the set, provided it is of a (homogeneous) sort and not a (heterogeneous) sum event. This excludes SVCs because the maximal event is not homogeneous. A CC is admissible because there is only one event in the set which is of a basic sort.

6

CONCLUSION

In this article, we presented an analysis of SVCs and CCs in the Kwa language Edo. The basic idea of our analysis is that SVCs and CCs denote complex event structures that are derived from simple ones denoted by verbs in isolation. At the semantic level verbs that occur as first verbs in one of these constructions are interpreted as mapping a VP denotation to an n-ary relation denoting a complex event structure. For SVCs, the events in this structure are linked either by a plan (CSVC) or a causal relation (RSVC). For a CC, the events are only related by temporal succession. SVCs are interpreted relative to the join consisting of the events in the sequence whereas a CC is interpreted relative to e_1 . As a result, manner adverbs modifying a complex predicate express a property of the complex event in an SVC and of e_1 in a CC. From this semantic characterization it follows that at the syntactic level (first) verbs in complex predicates take an additional argument of type vp . Hence, SVCs and CCs express complex event structures without using overt coordination or subordination. The application of structural rules like permutation and contraction at the syntactic

level is enforced by a combination of modal decoration and K-rules. Modal decorations are used for verbs and NPs though the way they are decorated is different.

We will close by mentioning two open questions and directions for future work. Since the use of a contraction rule does not guarantee the finite reading property, it is interesting to look for an alternative analysis which dispenses with such a rule. A second question concerns the analysis of CCs in which the subject of V_2 , which is coreferential with the subject of V_1 , is realized by an overt pronoun. The analysis presented in this article does not capture this case but only those in which this subject is not overtly realized. Furthermore, the analysis must be extended to negated and other types of adverbially modified multiverb sequences. Due to lack of space, no analysis of manner adverbs could be given.

APPENDIX:
MULTIMODAL NON-ASSOCIATIVE
LAMBEK-CALCULUS
WITH UNARY MULTIPLICATIVE OPERATORS

The base logic from the landscape of substructural logics that is used in this article is a multimodal variant of the non-associative Lambek calculus enriched with unary (modal) operators (or connectives) that function as control devices. This logic will be referred to by $NL(\diamond)$. We start by defining the categorial language. A *categorial formula* (or *category*) is inductively defined on the basis of a set Ω of atomic category formulas and a set $I \in I$ as

$$\Phi ::= \Omega \mid \Phi /_i \Phi \mid \Phi \bullet_i \Phi \mid \Phi \backslash_i \Phi \mid \diamond \Phi \mid \square \Phi$$

The collection of categorial formulas, inductively defined on the basis of Ω and I , will also be referred to by $CAT_I(\Omega)$. For the fragment of Edo considered in this article, it is sufficient to set $\Omega = \{np, s\}$. The elements of I are modes of compositions. Each family $\{/_i, \bullet_i, \backslash_i\}$ is interpreted relative to a ternary accessibility relation R_i . By contrast, the unary connectives are interpreted relative to a binary accessibility

relation R_\diamond . Given a valuation ν that assigns to each atomic categorial formula a subset of a set W of linguistic resources, ν is extended to complex formulas as given in (96).²⁵

- (96) a. $\nu(A \bullet_i B) = \{x \mid \exists y \exists z [R_i(x, y, z) \wedge y \in \nu(A) \wedge z \in \nu(B)]\}$
 b. $\nu(C /_i B) = \{y \mid \forall x \forall z [(R_i(x, y, z) \wedge z \in \nu(B)) \rightarrow x \in \nu(C)]\}$
 c. $\nu(A \setminus_i B) = \{z \mid \forall x \forall y [(R_i(x, y, z) \wedge y \in \nu(A)) \rightarrow x \in \nu(B)]\}$
 d. $\nu(\diamond A) = \{x \mid \exists y [R_\diamond(x, y) \wedge y \in \nu(A)]\}$
 e. $\nu(\square A) = \{x \mid \forall y [R_\diamond(y, x) \rightarrow y \in \nu(A)]\}$

The set Σ of antecedent terms (or structures) is inductively defined by $\Sigma ::= \Omega \mid (\Sigma \circ_i \Sigma) \mid \langle \Sigma \rangle$. The binary structural connectives \circ_i match the \bullet_i at the level of categorial formulas. Analogously, $\langle \cdot \rangle$ matches the unary connective \diamond .²⁶

The relation between syntax and semantics is based on a function $\tau : \text{CAT}_T(\Omega) \mapsto \text{Types}$. The set of types is defined below.

DEFINITION 2 *Types* *The set of basic types is $\text{Base} = \{e, t\}$. Given Base , the set of types Types is the smallest set s.t.*

- $\text{Base} \subseteq \text{Types}$,
- $\langle a, b \rangle \in \text{Types}$, if $a \in \text{Types}$ and $b \in \text{Types}$.

The mapping τ from syntactic types to semantic types is driven by the semantic interpretation of SVCs and CCs. Since we are working in a Neo-Davidsonian event framework, verbs in general get an additional (last) argument of sort ‘event’. This has the effect that after discharging the $n - 1$ non-event arguments one gets a term of type $\langle e, t \rangle$, i.e. a set of events. Standardly, one gets a term of type t by applying existential closure ($\lambda P. \exists e. P(e)$). We will not implement this operation and assume that the syntactic type s is mapped to the semantic type $\langle e, t \rangle$: $\tau(s) = \langle e, t \rangle$.²⁷ Since we do not treat quantification, the syntactic type np is mapped to the semantic type e : $\tau(np) = e$.

²⁵ Thus, categorial formulas are interpreted relative to frames $\langle W, \{R_i\}_{i \in I}, R_\diamond \rangle$.

²⁶ Instead of \circ_i and $\langle \cdot \rangle$ one also finds $(\cdot)^i$ and $(\cdot)^\diamond$. Thus, one has $(\Sigma, \Sigma)^i$ and $(\Sigma)^\diamond$.

²⁷ See Winter and Zwarts (2011) for one way of how such an operation can be incorporated into (abstract) categorial grammar. Our mapping for s resembles that in possible world semantics where sentences are propositions, i.e. sets of possible worlds.

- (97) a. $\tau(np) = e$.
 b. $\tau(s) = \langle e, t \rangle$.
 c. $\tau(A \setminus_i B) = \tau(A /_i B) = \langle \tau(A), \tau(B) \rangle$.
 d. $\tau(A \bullet B) = \tau(A) \times \tau(B)$.

Unary modalities are semantically inactive so that one has $\tau(\Box A) = \tau(\Diamond A) = \tau(A)$, Morrill (1994).

Given the mapping τ , each category formula (syntactic type) A assigned to a lexical item is paired with a typed λ -term representing the meaning of the item when it is assigned the syntactic type A . The set of λ -terms is defined below.

DEFINITION 3 Typed λ -term VAR_α is a countable infinite set of variables of type α and CON_α a set of constants of type α . The set $\lambda\text{-term}_\alpha$ of types λ -terms of type α is recursively defined as:

- $VAR_\alpha \subseteq \lambda\text{-term}_\alpha$,
- $CON_\alpha \subseteq \lambda\text{-term}_\alpha$,
- $t(t') \in \lambda\text{-term}_\beta$ if $t \in \lambda\text{-term}_{\langle \alpha, \beta \rangle}$ and $t' \in \lambda\text{-term}_\alpha$,
- $\lambda x.t \in \lambda\text{-term}_{\langle \alpha, \beta \rangle}$ if $x \in VAR_\alpha$ and $t \in \lambda\text{-term}_\beta$.

$Term = \bigcup_{\alpha \in \text{Types}} \lambda\text{-term}_\alpha$ is the set of all (typed) λ -terms. Given a model \mathfrak{M} and a variable assignment θ , the denotation (or interpretation) of a λ -term is defined as follows: (i) $\llbracket x \rrbracket_{\mathfrak{M}}^\theta = \theta(x)$ if $x \in VAR_\alpha$, (ii) $\llbracket c \rrbracket_{\mathfrak{M}}^\theta = \llbracket c \rrbracket$ if $c \in CON_\alpha$, (iii) $\llbracket t(t') \rrbracket_{\mathfrak{M}}^\theta = \llbracket t \rrbracket_{\mathfrak{M}}^\theta(\llbracket t' \rrbracket_{\mathfrak{M}}^\theta)$, and (iv) $\llbracket \lambda x.t \rrbracket_{\mathfrak{M}}^\theta = f$ such that $f(a) = \llbracket t \rrbracket_{\mathfrak{M}}^{\theta[x:=a]}$.

Sequents are annotated with λ -terms. A sequent is a pair (Γ', B') . Γ' is of the form $(x_1 : A_1, \dots, x_n : A_n)$ where each $A_i \in \Sigma$ and the variables x_i in the antecedent are mutually distinct. B' is of the form $t : B$ where $B \in \Phi$ and the term t is constructed out of the x_i . Hence, a derivation of an annotated sequent represents the computation of a denotation recipe t of (syntactic) type B with input parameters x_i of (syntactic) type A_i , Moortgat (1997). Sequents are written as $\Gamma \Rightarrow B$.

The logic is a combination of inference rules for the constructors $\{/, \setminus, \bullet, \Box, \Diamond\}$, relativized to a particular mode, and a set of structural rules of inference for the manipulation of the antecedents in sequents. Below, a sequent presentation of $NL(\Diamond)$ in the Natural Deduction format is given. Besides the identity axiom and the cut rule (which is

eliminable), one has as inference rules introduction and elimination rules for each binary and unary connective.

The base logic $NL(\diamond)$:

$$[Ax] \frac{}{x : A \Rightarrow x : A}$$

$$[/_iI] \frac{(\Gamma \circ_i x : B) \Rightarrow t : A}{\Gamma \Rightarrow \lambda x.t : A/_iB}$$

$$[\backslash_iI] \frac{(x : B \circ_i \Gamma) \Rightarrow t : A}{\Gamma \Rightarrow \lambda x.t : B \backslash_i A}$$

$$[\bullet_iI] \frac{\Gamma \Rightarrow t : A \quad \Delta \Rightarrow u : B}{(\Gamma \circ_i \Delta) \Rightarrow \langle t, u \rangle : A \bullet_i B}$$

$$[\Box I] \frac{\langle \Gamma \rangle \Rightarrow t : A}{\Gamma \Rightarrow t : \Box A}$$

$$[\Diamond I] \frac{\Gamma \Rightarrow t : A}{\langle \Gamma \rangle \Rightarrow t : \Diamond A}$$

$$\frac{\Gamma \Rightarrow t : A \quad \Delta[x : A] \Rightarrow u : C}{\Delta[\Gamma] \Rightarrow u[t/x] : C} [Cut]$$

$$\frac{\Gamma \Rightarrow t : A/_iB \quad \Delta \Rightarrow u : B}{(\Gamma \circ_i \Delta) \Rightarrow (tu) : A} [/_iE]$$

$$\frac{\Gamma \Rightarrow u : B \quad \Delta \Rightarrow t : B \backslash_i A}{(\Gamma \circ_i \Delta) \Rightarrow (tu) : A} [\backslash_iE]$$

$$\frac{\Delta \Rightarrow u : A \bullet_i B \quad \Gamma[x : A \circ_i y : B] \Rightarrow t : C}{\Gamma[\Delta] \Rightarrow t[\pi^0(u)/x, \pi^1(u)/x] : C} [\bullet_iE]$$

$$\frac{\Gamma \Rightarrow t : \Box A}{\langle \Gamma \rangle \Rightarrow t : A} [\Box E]$$

$$\frac{\Delta \Rightarrow u : \Diamond A \quad \Gamma[\langle x : A \rangle] \Rightarrow t : B}{\Gamma[\Delta] \Rightarrow t[u/x] : B} [\Diamond E] .$$

Whereas the logical rules are fixed, the structural rules depend on the application. Since we are using a multimodal setting, the structural rules are relativized to particular modes. The following modes of composition are distinguished for Edo.

- \cdot_{1r} : right-headed verb-complement (subject-verb relation)
- \cdot_{1l} : left-headed verb-complement (non-subject (object)-verb relation)
- \cdot_0 : verb-adjunction mode for an CSVC (relation between extended verb and additional argument in this kind of SVC)
- \cdot_2 : verb-adjunction mode for an RSVC and a CC (relation between extended verb and additional argument in these two kinds of multiverb sequences)
- \cdot_{rd} : head wrapping mode for ditransitive verbs

Thus, in the present context $I = \{\cdot_{1r}, \cdot_{1l}, \cdot_0, \cdot_2, \cdot_{rd}\}$. Given I , $NL(\diamond)$ is extended by the following structural rules. As already said above, this kind of rule is used to manipulate the antecedents of sequents. Furthermore, except for the rule of contraction, structural rules are semantically inert, i.e. they do not operate on the λ -term in the consequent. We give both the algebraic and the natural deduction sequent presentation.²⁸

K-Rules:

$$\text{a. } K(\bullet_{1r}): \diamond(A \bullet_{1r} B) \rightarrow \diamond A \bullet_{1r} \diamond B$$

$$\frac{\Gamma[(\langle \Delta \rangle \circ_{1r} \langle \Delta' \rangle)] \Rightarrow t : C}{\Gamma[(\langle \Delta \rangle \circ_{1r} \Delta')] \Rightarrow t : C} [K(\bullet_{1r})]$$

$$\text{b. } K^*2(\bullet_{1l}): \diamond(\diamond A \bullet_{1l} B) \rightarrow \diamond A \bullet_{1l} \diamond B$$

$$\frac{\Gamma[(\langle \Delta \rangle \circ_{1l} \langle \Delta' \rangle)] \Rightarrow t : C}{\Gamma[(\langle \langle \Delta \rangle \rangle \circ_{1l} \Delta')] \Rightarrow t : C} [K^*2(\bullet_{1l})]$$

²⁸ Assuming that structural rules are formulated using only the unary connective \diamond and the \bullet_i from the logical vocabulary of the categorial language, there is the following back-and-forth translation between the two representations. A rule $A \rightarrow B$ in the algebraic format corresponds to a rule of inference that admits to replace a subterm Δ' in the premise by Δ in the conclusion, with Δ and Δ' the equivalences of A and B , respectively:

$$A \rightarrow B \rightsquigarrow \frac{\Gamma[\Delta'] \Rightarrow t : C}{\Gamma[\Delta] \Rightarrow t : C} .$$

- c. $K1(\bullet_0): \diamond(A \bullet_0 B) \rightarrow \diamond A \bullet_0 B$

$$\frac{\Gamma[(\langle \Delta \rangle \circ_0 \Delta')] \Rightarrow t : C}{\Gamma[(\langle \Delta \circ_0 \Delta' \rangle)] \Rightarrow t : C} [K1(\bullet_0)]$$
- d. $K(\bullet_2): \diamond(A \bullet_2 B) \rightarrow \diamond A \bullet_2 \diamond B$

$$\frac{\Gamma[(\langle \Delta \rangle \circ_2 \langle \Delta' \rangle)] \Rightarrow t : C}{\Gamma[(\langle \Delta \circ_2 \Delta' \rangle)] \Rightarrow t : C} [K(\bullet_2)]$$
- e. $K(l/rd): A \bullet_{1l} B \rightarrow A \bullet_{rd} B$

$$\frac{\Gamma[(\Delta \circ_{rd} \Delta')] \Rightarrow t : C}{\Gamma[(\Delta \circ_{1l} \Delta')] \Rightarrow t : C} [K(l/rd)]$$
- f. $K^*(\bullet_{1l}): \diamond(\langle \diamond A \bullet_{rd} B \rangle \bullet_{1l} C) \rightarrow \diamond(\diamond A \bullet_{rd} B) \bullet_{1l} \diamond C$

$$\frac{\Gamma[(\langle \langle \Delta \rangle \circ_{rd} \Delta' \rangle \circ_{1l} \langle \Delta'' \rangle)] \Rightarrow t : C}{\Gamma[(\langle \langle \Delta \rangle \circ_{rd} \Delta' \rangle \circ_{1l} \Delta'')] \Rightarrow t : C} [K^*(\bullet_{1l})]$$
- g. $K^*2(\bullet_{rd}): \diamond(\diamond A \bullet_{rd} B) \rightarrow \diamond A \bullet_{rd} \diamond B$

$$\frac{\Gamma[(\langle \Delta \rangle \circ_{rd} \langle \Delta' \rangle)] \Rightarrow t : C}{\Gamma[(\langle \langle \Delta \rangle \circ_{rd} \Delta' \rangle)] \Rightarrow t : C} [K^*2(\bullet_{rd})]$$
- h. $K^*(\bullet_0): \diamond(A \bullet_0 (\diamond B \bullet_{rd} C)) \rightarrow \diamond A \bullet_0 \diamond(\diamond B \bullet_{rd} C)$

$$\frac{\Gamma[(\langle \Delta \rangle \circ_0 \langle \langle \Delta' \rangle \circ_{rd} \Delta'' \rangle)] \Rightarrow t : C}{\Gamma[(\langle \Delta \circ_0 (\langle \Delta' \rangle \circ_{rd} \Delta'') \rangle)] \Rightarrow t : C} [K^*(\bullet_0)]$$

Mixed Permutation Rules:

- a. $MP1: (A \bullet_{1l} \diamond B) \bullet_i C \rightarrow (A \bullet_i C) \bullet_{1l} \diamond B \quad i = 0 \text{ or } i = 2$

$$\frac{\Gamma[(\langle \langle \Delta \circ_i \Delta'' \rangle \circ_{1l} \langle \Delta' \rangle \rangle)] \Rightarrow t : C}{\Gamma[(\langle \langle \Delta \circ_{1l} \langle \Delta' \rangle \rangle \circ_i \Delta'' \rangle)] \Rightarrow t : C} [MP1]$$
- b. $MP2: (A \bullet_{rd} B) \bullet_{1l} C \rightarrow (A \bullet_{1l} C) \bullet_{rd} B$

$$\frac{\Gamma[(\langle \langle \Delta \circ_{1l} \Delta'' \rangle \circ_{rd} \Delta' \rangle)] \Rightarrow t : C}{\Gamma[(\langle \langle \Delta \circ_{rd} \Delta' \rangle \circ_{1l} \Delta'' \rangle)] \Rightarrow t : C} [MP2]$$
- c. $MP3: (A \bullet_{rd} B) \bullet_0 C \rightarrow (A \bullet_0 C) \bullet_{rd} B$

$$\frac{\Gamma[(\langle \langle \Delta \circ_0 \Delta'' \rangle \circ_{rd} \Delta' \rangle)] \Rightarrow t : C}{\Gamma[(\langle \langle \Delta \circ_{rd} \Delta' \rangle \circ_0 \Delta'' \rangle)] \Rightarrow t : C} [MP3]$$

Mixed Contraction Rule:

$$\text{a. MC: } (A \bullet_0 B) \bullet_{1l} \diamond C \rightarrow (A \bullet_{1l} \diamond C) \bullet_0 (B \bullet_{1l} \diamond C)$$

$$\frac{\Gamma[(\Delta_1 \circ_{1l} \langle x : \Delta_3 \rangle) \circ_0 (\Delta_2 \circ_{1l} \langle y : \Delta_3 \rangle)] \Rightarrow t : C}{\Gamma[(\Delta_1 \circ_0 \Delta_2) \circ_{1l} \langle x : \Delta_3 \rangle] \Rightarrow t[y \leftarrow x] : C} \text{ [MC]}$$

The types vp and tv are defined in the usual way.

$$(98) \quad \begin{array}{l} \text{a. } \text{vp} =_{\text{def.}} \text{np} \backslash_{1r} \text{s} \\ \text{b. } \text{tv} =_{\text{def.}} \text{vp} /_{1l} \text{np} \end{array}$$

Let Ψ be the set of structural rules given above. The logic to be used in the sections to follow is $\text{NL}(\diamond)$ plus the structural rules in Ψ . This logic will be referred to as $\text{NL}(\diamond) + \Psi$. The notion of *Lambek Grammar* is defined as follows.²⁹

DEFINITION 4 Lambek Grammar *Let Θ be an alphabet. A Lambek grammar G is a triple (Ω, LEX, S) , where Ω is a finite set (i.e. the set of basic categorial formulas), LEX is a finite subrelation of $\Theta^+ \times \text{CAT}_I(\Omega)$ (with an index set I), and S is a finite subset of $\text{CAT}_I(\Omega)$ (the designated categorial formulas).*

For Edo, the designated categorial formula is \square s. This is empirically motivated in Section 5.1. In the presence of a semantic component, one gets a term-labeled lexicon. $\text{LEX} \subseteq \Theta^+ \times (\text{CAT}_I(\Omega) \times \text{Term})$. One has: if $\langle w, \langle A, t \rangle \rangle \in \text{LEX}$ then $t \in \lambda\text{-term}_{\tau(A)}$.

A Lambek grammar G determines a language over Θ in the following way.³⁰

DEFINITION 5 Language determined by a Lambek Grammar *Let $G = \langle \Omega, \text{LEX}, S \rangle$ be a Lambek grammar over the alphabet Θ . Then $\alpha \in L(G)$ iff there are $a_1, \dots, a_n \in \Theta^+$, $(A_1, \dots, A_n) \in \text{CAT}_I(\Omega)$, and $S \in S$ such that*

- (i) $\alpha = a_1, \dots, a_n$
- (ii) for all i such that $1 \leq i \leq n$: $\langle a_i, A_i \rangle \in \text{LEX}$, and
- (iii) $\text{NL}(\diamond) + \Psi \vdash (A_1, \dots, A_n) \Rightarrow S$.

²⁹See Jäger (2005) for details from which the following definitions are adapted.

³⁰Note that the lexicon is defined without reference to the Curry-Howard correspondence. The adaption of the definition to labeled sequents is straightforward.

In Definition 5, \vdash is the relation of derivability relative to $NL(\diamond)+\Psi$. (A_1, \dots, A_n) is a binary bracketed structure. If for a sequent $(A_1, \dots, A_n) \Rightarrow S$ such that $NL(\diamond)+\Psi \vdash (A_1, \dots, A_n) \Rightarrow S \in \mathcal{S}$ there is a sequence $\alpha = a_1, \dots, a_n$ such that for all i with $1 \leq i \leq n$: $\langle a_i, A_i \rangle \in \text{LEX}$, the sequent $(A_1, \dots, A_n) \Rightarrow S$ is said to admit of a *lexical substitution*, meaning that the sequent is an element of $L(G)$, i.e. the language determined by G . Basing the definition of terms (or structures) Σ not only on the set Ω of categorial formulas but also on the subset of Θ^+ consisting of those elements occurring in the domain of LEX (i.e. the set $\{a \in \Theta^+ \mid \text{there is an } A \text{ in } \text{CAT}_I(\Omega) \text{ s.t. } \langle a, A \rangle \in \text{LEX}\} = \text{dom}(\text{LEX})$), an element $\langle a, A \rangle \in \text{LEX}$ can be taken as a *lexical axiom*, written $a \Rightarrow A$.

The way modalities are used in this article was first introduced in Moortgat (1996) and extended in Moortgat (1997) and Kurtonina (1995). Kurtonina and Moortgat (1997) develop a theory of communication between categorial type logics. It is shown how one can recover the structural discrimination of a weaker logic from within a stronger one (structural inhibition) and how one can reintroduce structural relaxation of stronger logics within weaker ones.

Monomodal NL is sound and complete with respect to the interpretation of unary and binary connectives given in (96) (see Moot and Retoré 2012 for a proof and details). For the multimodal variant, the situation is more complicated (see again Moot and Retoré 2012 for details and references cited therein). NL is strictly context-free and has a polynomial recognition problem. The move to a multimodal variant without structural rules does not lead beyond context-free recognition. The relation between multimodality, structural rules and unary modalities is more complicated. If no copying and deletion are allowed for structural rules and if the unary modalities are non-expanding, one obtains the full expressivity of context-sensitive grammars, and the PSPACE complexity that goes with it. If no restrictions are imposed on structural rules (specifically, if one allows copying and deletion operations), one obtains the expressivity of unrestricted rewriting systems.

ACKNOWLEDGEMENTS

This work was supported by the CRC991 “The Structure of Representations in Language, Cognition, and Science” funded by the German Research Foundation (DFG). We are grateful to the reviewers of this paper for their valuable comments and suggestions.

REFERENCES

- Rebecca AGHEYISI (1986), *Edo-English dictionary*, Ethiope Publishers, Benin City.
- Alexandra AIKHENVALD (2006), Serial verb constructions, in Alexandra AIKHENVALD and Robert Malcolm Ward DIXON, editors, *Serial verb constructions in typological perspective*, pp. 1–68, Oxford University Press.
- Nicholas ASHER and Alex LASCARIDES (2001), *Logics of conversation*, Cambridge University Press.
- Mark BAKER and Osamuyimen Thompson STEWART (1999), On double headedness and the anatomy of the clause, ms. Rutgers University.
- Mark BAKER and Osamuyimen Thompson STEWART (2001), A serial verb construction without constructions, ms. Rutgers University.
- Jürgen BOHNEMEYER, Nicholas J. ENFIELD, James ESSEGBEY, Iraide IBARRETXE-ANTUÑANO, Sotaro KITA, Friederike LÜPKE, and Felix K. AMEKA (2007), Principles of event segmentation in language: The case of motion events, *Language*, 83(3):495–532.
- Oliver BOTT and Wolfgang STERNEFELD (2017), An event semantics with continuations for incremental interpretation, *Journal of Semantics*, 34(2):201–236.
- Lucas CHAMPOLLION (2015), The interaction of compositional semantics and event semantics, *Linguistics and Philosophy*, 38(1):31–66.
- Robert Malcolm Ward DIXON (2006), Serial verb constructions: conspectus and coda, in Alexandra AIKHENVALD and Robert Malcolm Ward DIXON, editors, *Serial verb constructions – a cross-linguistic typology*, pp. 338–350, Oxford University Press.
- William FOLEY (1991), *The Yimas language of New Guinea*, Stanford University Press.

- William A. FOLEY (2010), Events and serial verb constructions, in Mengistu AMBERBER, Brett BAKER, and Mark HARVEY, editors, *Complex predicates: cross-linguistic perspectives on event structure*, pp. 79–109, Cambridge University Press.
- Thomas GAMERSCHLAG (2005), *Komposition und Argumentstruktur komplexer Verben. Eine lexikalische Analyse von Verb-Verb-Komposita und Serialverbkonstruktionen*, volume 61 of *Studia grammatica*, Akademie Verlag.
- Jane GRIMSHAW (1990), *Argument structure*, MIT Press.
- Ken HALE and Samuel J. KEYSER (1993), On argument structure and the lexical expression of syntactic relations, in Ken HALE and Samuel J. KEYSER, editors, *The view from building 20*, pp. 53–109, MIT Press.
- Lars HELLAN, Dorothee BEERMANN, and Eli SÆTHERØ ANDENES (2003), Towards a typology of serial verb constructions in Akan, in Mary Esther KROPP DAKUBU and Kweku E. OSAM, editors, *Proceedings of the annual colloquium of the Legon-Trondheim linguistics project*, Studies in the languages of the Volta basin, pp. 61–86.
- Gerhard JÄGER (2005), *Anaphora and Type Logical Grammar*, Trends in Logic 24, Springer.
- Angelika KRATZER (1996), Severing the external argument from its verb, in Johan ROORYCK and Laurie ZARING, editors, *Phrase Structure and the Lexicon*, pp. 109–138, Kluwer.
- Natasha KURTONINA (1995), *Frames and labels*, Phd., OTS, University of Utrecht & ILLC, University of Amsterdam.
- Natasha KURTONINA and Michael MOORTGAT (1997), Structural control, in Patrick BLACKBURN and Maarten DE RIJKE, editors, *Specifying syntactic structures*, Studies in Language, Logic and Information, pp. 75–114, CSLI Press.
- Godehard LINK (1998), *Algebraic Semantics in Language and Philosophy*, CSLI Publications, Stanford.
- Michael MOORTGAT (1996), Multimodal linguistic inference, *Journal of Logic, Language and Information*, 5(3):349–385.
- Michael MOORTGAT (1997), Categorical type logics, in Johan van BENTHEM and Alice ter MEULEN, editors, *Handbook of logic and language*, pp. 93–177, North-Holland.
- Michael MOORTGAT and Richard OERHLE (1993), Adjacency, dependency and order, in *Proceedings of the 9th Amsterdam Colloquium*, pp. 447–466.
- Richard MOOT and Christian RETORÉ (2012), *The logic of categorial grammars: a deductive account of natural language syntax and semantics*, Springer.
- Glyn MORRILL (1994), *Type Logical Grammar*, Kluwer.

A type-logical analysis of SVCs and CCs in Edo

Glyn MORRILL (1995), Discontinuity in categorial grammar, *Linguistics and Philosophy*, 18(2):175–219.

Glyn MORRILL (2011), *Categorial grammar: logical syntax, semantics, and processing*, Oxford University Press.

Ralf NAUMANN and Wiebke PETERSEN (2019), Bridging inferences in a dynamic frame theory, in Alexandra SILVA, Sam STATON, Peter SUTTON, and Carla UMBACH, editors, *Language, logic, and computation*, pp. 228–252, Springer.

Ota OGIE (2010), *Multi-verb constructions in Edo*, VDM Verlag.

Osamuyimen Thompson STEWART (1996), Adverb placement and the structure of the serial verb construction, in *Proceedings of the North East Linguistic Society* 26, pp. 409–423.

Osamuyimen Thompson STEWART (2001), *The serial verb construction parameter*, Garland.

Yoad WINTER and Joost ZWARTS (2011), Event semantics and abstract categorial grammar, in Makoto KANAZAWA, András KORNAI, Marcus KRACHT, and Hiroyuki SEKI, editors, *The mathematics of language*, pp. 174–191, Springer.

Ralf Naumann

© 0000-0002-0222-5540

naumann@phil-fak.uni-duesseldorf.de

Thomas Gamerschlag

© 0000-0001-7996-9137

gamer@phil-fak.uni-duesseldorf.de

Heinrich-Heine-Universität Düsseldorf
Düsseldorf, Germany

Ralf Naumann and Thomas Gamerschlag (2020), *Serial verb constructions and covert coordinations in Edo – an analysis in Type Logical Grammar*, *Journal of Language Modelling*, 8(2):337–413

doi <https://dx.doi.org/10.15398/jlm.v8i2.221>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

© <http://creativecommons.org/licenses/by/4.0/>

A French corpus annotated for multiword expressions and named entities

Marie Candito¹, Mathieu Constant², Carlos Ramisch³, Agata Savary⁴,
Bruno Guillaume⁵, Yannick Parmentier^{6,7} and Silvio Ricardo Cordeiro¹

¹Université de Paris, CNRS, LLF

² Université de Lorraine, CNRS, ATILF

³ Aix Marseille Univ, Université de Toulon, CNRS, LIS

⁴Université de Tours, LIFAT

⁵Université de Lorraine, CNRS, Inria, LORIA

⁶Université de Lorraine, CNRS, LORIA

⁷Université d'Orléans, LIFO

ABSTRACT

We present the enrichment of a French treebank of various genres with a new annotation layer for multiword expressions (MWEs) and named entities (NEs).¹ Our contribution with respect to previous work on NE and MWE annotation is the particular care taken to use formal criteria, organized into decision flowcharts, shedding some light on the interactions between NEs and MWEs. Moreover, in order to cope with the well-known difficulty to draw a clear-cut frontier between compositional expressions and MWEs, we chose to use sufficient criteria only. As a result, annotated MWEs satisfy a varying number of sufficient criteria, accounting for the scalar nature of the MWE status. In addition to the span of the elements, annotation includes the subcategory

Keywords:
multiword
expressions,
annotation,
corpus, French

¹For verbal MWEs, we have reused the annotation performed within the PARSEME COST multilingual project (Savary *et al.* 2017), so the present article focuses on named entities and non-verbal MWEs.

of NEs (e.g., person, location) and one matching sufficient criterion for non-verbal MWEs (e.g., lexical substitution). The 3,099 sentences of the treebank were double-annotated and adjudicated, and we paid attention to cross-type consistency and compatibility with the syntactic layer. Overall inter-annotator agreement on non-verbal MWEs and NEs reached 71.1%. The released corpus contains 3,112 annotated NEs and 3,440 MWEs, and is distributed under an open license.

1

INTRODUCTION

Multiword expressions (MWEs) such as idioms (e.g., *dead end*, *break the ice*) and light-verb constructions (e.g., *make decision*) have been the focus of a vast amount of linguistic studies and annotation projects (reviewed in Section 2). The idiosyncrasy at the heart of the concept of MWE is a challenge for any linguistic theory and disrupts automatic processing, as MWEs mix idiosyncratic and regular patterns. Because of their partly unpredictable behavior, MWEs have been widely listed in lexicons and annotated in corpora. Yet, for many languages, MWE-annotated resources are generally not associated with operational decision criteria, the guidelines being often reduced to examples of the various MWE categories.

Corpora annotated for named entities (NEs) such as person (e.g., *Theresa May*) and location (e.g., *Colombia*) also abound in many languages.² However, the overlap between MWEs and NEs has rarely been studied. Given these challenges, our first objective is to provide operational criteria for defining MWEs on the one hand and NEs on the other hand, so that both categories can be precisely distinguished and annotated within the same framework. Secondly, we test the proposed criteria against actual annotation in a French corpus. We chose not to use pre-existing MWE and NE lexicons, to avoid biases, but we use post-annotation coherence checking tools to improve cross-type consistency of annotations.

²Our work covers single-word and multiword NEs. Although multiword NEs can be considered MWEs, hereafter we reserve the term MWE for expressions that are not NEs, see Section 3.2 for details.

A fundamental trait of our approach is to model the MWE status in parallel to the syntactic layer: depending on its distribution and internal pattern, a given MWE can be considered syntactically regular, hence receiving a regular internal structure. Another originality stands in our choice to use sufficient criteria for the MWE status, in order to cope with their varying degree of idiosyncrasy. Indeed, when applied to non-prototypical MWE examples, MWE criteria may often contradict each other. We thus opted for sufficient criteria, instead of relying on a subjective quantification of how many and which criteria should prevail. The resulting resource thus comprises annotated MWEs with varying degrees of idiosyncrasy.

The remainder of this article is organized as follows: in Section 2 we discuss related work, covering the general MWE definition and typologies, their annotation in corpora, and NE annotation. In Section 3, we present and motivate the main distinctions we made, in particular between NEs and MWEs, and present our typologies. Section 4 describes the formal constraints for our MWEs and NEs, and the top decision flowchart guiding the annotators to the various sub-guides. Section 5 is devoted to the guidelines for NEs, Section 6 summarizes the guidelines for verbal MWEs defined in the PARSEME project, and Section 7 describes our guidelines for non-verbal MWEs. In Section 8 we describe the source corpus, the annotation process and annotation quality. Section 9 is devoted to the interaction between MWEs and syntactic annotations. Finally, we present various statistics for the resulting resource in Section 10, we mention some lessons learned from the project in Section 11 and we conclude in Section 12.

RELATED WORK

2

This section presents some of the previous work in the field of MWE and NE annotation. Due to their extensive use in multiple information extraction tasks, NEs have received by far much more attention than MWEs in the last two decades. We have thus decided to put a stronger emphasis on prior work in MWE annotation. We first provide various definitions for the term “multiword expression” that encompasses a wide body of linguistic phenomena (Section 2.1). Then, we summarize

existing MWE typologies (Section 2.2). Next, we present emblematic initiatives for MWE-annotated corpora and treebanks, focusing on the criteria and tests used (Section 2.3). Finally, we synthesize the large body of work on NE annotation in corpora (Section 2.4).

2.1

MWE definitions

The term *multiword expression* (MWE) has emerged in the natural language processing (NLP) community in the early 2000s, notably in the famous paper of Sag *et al.* (2002). The authors roughly define MWEs as “idiosyncratic interpretations that cross word boundaries (or spaces)”, emphasizing the unpredictability of their linguistic behavior. This informal definition actually captures a wide body of heterogeneous linguistic phenomena, including phrasal verbs, idioms, light-verb constructions, complex function words, and nominal compounds. Since then, many other definitions have been proposed (Constant *et al.* 2017). Among others, Baldwin and Kim (2010) propose a more precise definition, stating that MWEs are “lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomatity”. They provide an overview of the main properties for every type of idiomatity, as well as a simple procedure to test whether a candidate word combination is an MWE or not, by testing all types of idiomatity. Still, this definition is not operational because it does not indicate the precise individual idiomatity tests to apply systematically. NLP researchers tend to give rough definitions of MWEs, and illustrate them with lists of categories and examples to specify the concept denoted by the term. These usually emphasize the idiosyncratic nature of these expressions, and the difficulty to process them from a computational (linguistic) point of view. There are several reasons for this vagueness.

First, the status of MWEs is not clearly defined from a linguistic point of view. As they are located at the lexicon-grammar interface, their definition depends on the underlying linguistic framework. MWEs are highly related to phraseology, a historical field of linguistics in which researchers have been extensively describing MWEs for several decades. Mel’čuk (2012) goes even further, stressing that “there is no agreement on either the exact content of the notion of ‘phraseology’, nor on the way phraseological expressions should be described,

nor on how they should be treated in linguistic applications, in particular, in lexicography and Natural Language Processing”.

Second, from an NLP point of view, MWEs embrace word combinations that need to be considered as units at some level of linguistic processing (Calzolari *et al.* 2002). As a consequence, in NLP, the set of considered MWEs heavily depends on the target application. For instance, Copestake *et al.* (2002) suggest that idiomatic expressions with regular syntactic structures are of no use in a system producing syntactic trees.³ Furthermore, NLP models heavily rely on linguistic resources, in particular MWE resources in case of MWE-aware models. The role of precisely defining MWEs is therefore entrusted to the resource designers. Indeed, building an MWE-aware resource requires a set of operational criteria to identify them: either to create and encode MWE entries in lexical resources, or to annotate them in corpora.

Formal criteria are especially useful to operationalize (vague) MWE definitions. Historically, formal criteria have been designed mainly for lexicographic purposes, on top of linguistic studies. Such criteria are usually based on the fact that the fixedness of one or several component(s) of a candidate MWE entails some idiomaticity. Fixedness is characterized by the fact that applying a transformation to a given MWE leads to unexpected meaning shifts or unacceptable sequences compared to similar linguistic contexts. For instance, the MWE *from time to time* does not accept modifier insertion (e.g., **from a time to another time*), whereas in similar linguistic contexts this is accepted (e.g., *from place to place* vs. *from a place to another place*). Gross (1986) applies formal criteria to classify and encode the properties of MWEs in a syntactic lexicon in French, the so-called *lexicon-grammar tables*.⁴ This formal approach largely inspired the guidelines used to annotate MWEs in various French corpora (Abeillé *et al.* 2003; Laporte *et al.* 2008b,a). It led to new definitions such as the one in Laporte *et al.* (2008b), who consider “a phrase composed of several words to be a multiword expression if some or all of their elements are frozen

³This claim, though very illustrative, has some counter-examples in the parsing literature: e.g. Cafferkey *et al.* (2007) show the positive impact of pre-identifying prepositional MWEs on syntactic constituency parsing accuracy.

⁴Lexicon-grammar tables have also been developed for other languages, e.g. Freckleton (1985) for English, and Català and Baptista (2007) for Spanish.

together in the sense of Gross (1986), that is, if their combination does not obey productive rules of syntactic and semantic compositionality”. In other words, we have a MWE if and only if its meaning cannot be derived from its individual components using a grammar including both a syntactic and a semantic component.

Recently, a breakthrough was witnessed in the way of defining MWEs with the creation of corpora annotated for verbal MWEs for the PARSEME shared tasks (Savary *et al.* 2017; Ramisch *et al.* 2018). The proposed definition is fully operational as it is entirely based on decision flowcharts relying on formal tests. Note that the main principles of this definition are in line with the ones adopted in our work. In our annotation of a French corpus with MWEs and NEs, we started by integrating the verbal MWE annotation of the French part of the PARSEME corpora (Section 6). Also note that, in the PARSEME annotation of verbal MWEs, as well as in our annotation of all kinds of MWEs, statistical idiomaticity (Baldwin and Kim 2010), that is, outstanding cooccurrence frequency, is not a sufficient criterion for the MWE status. Thus, “collocations” that do not satisfy other criteria are considered MWEs neither in PARSEME, nor in the present work.

2.2

MWE typologies

Because MWEs encompass heterogeneous linguistic objects, their description is usually accompanied by defining a typology of MWEs. Savary *et al.* (2018) present a comparison of several NLP-dedicated MWE typologies – those which were particularly influential, have been tested against representative datasets, or focus on verbal MWEs – proposed by Sag *et al.* (2002), Baldwin and Kim (2010), Mel’čuk (2010), Schneider *et al.* (2014), Laporte (2018), Sheinflux *et al.* (2019), and Savary *et al.* (2018) themselves. The analysis shows a large heterogeneity of these typologies in terms of:

- the number of languages covered – the first 6 works focus on a single language among English, French, and Hebrew whereas the last one covers 18 languages;
- the scope – from verbal MWEs only, to all syntactic categories of MWEs, including or not some categories of collocations;

- the number and granularity of MWE categories – from flat lists of 2–3 categories, to a 2–4-level hierarchy with 6–8 leaf categories;
- the number of classified expressions – from 15 to dozens of thousands of MWE lexicon entries or corpus occurrences;
- the criteria used for defining the categories – lexical (lexicalization, selection constraints, association strength), morphosyntactic (structure, presence of support verb, morphological and syntactic flexibility), semantic (decomposability, non-compositionality, transparency, figuration), and cross-lingual (universality).

Some works performed on French are inspired by the Meaning-Text Theory applied to phraseology by Mel'čuk (2010). For instance, Lux-Pogodalla and Polguère (2011), Polguère (2014) and Pausé (2017) integrate 4,400 collocations and 3,200 idioms in the French Lexical Network, where simple-word and multiword lexemes are densely interconnected. Mel'čuk's typology also inspired corpus annotation efforts by Tutin and Esperança-Rodier (2019), who notably extended it with multiword NEs and complex terms. They also defined a separate category for functional MWEs (adverbs, prepositions, conjunctions, determiners and pronouns). Let us finally mention the updated version of the PARSEME typology (Ramisch *et al.* 2018), with 5 main categories, 4 of which are relevant to French (Section 6).

MWE-annotated corpora and treebanks

2.3

We present some emblematic corpora annotated for MWEs, focusing on their annotation process and guidelines. We discuss annotation in syntactically non-annotated corpora (Section 2.3.1); and then in treebanks, in interaction with syntactic annotation (Section 2.3.2).

MWE annotation in corpora

2.3.1

Laporte *et al.* (2008b) and Laporte *et al.* (2008a) present the annotation process of a French corpus for adverbial and nominal compounds. The corpus (a Jules Verne's novel and parliamentary debates) contains 8,794 sentences, 168,856 words, 4,383 occurrences of MWEs with adverbial function, and 5,054 occurrences of multiword nouns. The annotation process starts with an automatic annotation based on compound dictionary lookup, followed by a manual validation

based on guidelines.⁵ These do not elaborate much on the linguistic tests/criteria to identify MWEs, but mainly rely on Gross (1986). For adverbial compounds, emphasis is laid on detecting when an MWE functions as an adverbial. Regarding multiword nouns, the guidelines focus on NEs and their category (place, person name, quotation), title and function nouns, nested MWEs, and non-predicating adjectives. The quality of the annotation process was not assessed.

Schneider *et al.* (2014) present a methodology for the annotation of a 55,000-word corpus of English web texts, the Streusle corpus. They aim at full coverage, with no limitations in terms of syntactic constructions, including both continuous and discontinuous MWEs. “Strong” and “weak” MWEs are distinguished, roughly corresponding to idiomatic MWEs and collocations. The guidelines are mainly a list of cases and examples (depending on the MWE structure). They rely on the following definition: MWEs are token combinations that are “idiosyncratic in form, function, or frequency”.⁶ The annotators’ judgements on the MWE status of a candidate expression are largely driven by their intuitions, informed by classical linguistic cues (e.g., semantic opacity, fixedness). Three types of annotation sessions were conducted: individual, joint and consensus sessions, with one, two, or more than two annotators collaborating. All sentences were annotated at least in one joint and one individual session, and 1/5 in a consensus session.

The Wiki50 corpus contains 50 English Wikipedia articles, totalling 4,350 sentences, annotated for NEs and MWEs (Vincze *et al.* 2011). A subset of 15 articles was double-annotated by linguists, and disagreements were discussed and resolved by the annotators themselves. The annotation scheme covers 6 MWE and 4 NE categories, with discontinuous expressions (light-verb and verb-particle constructions) represented using two-level hierarchical encoding. The MWE categories do not cover fixed adverbials nor functional MWEs, whereas the NE categories cover mainly person, organization and location. The

⁵ <http://infolingu.univ-mlv.fr/corpus/fr-MW-N/fr-MW-N/guidelines.doc> for nouns and <http://infolingu.univ-mlv.fr/corpus/fr-MW-Adv/fr-MW-Adv-corpus/guidelines.doc> for adverbials.

⁶ <https://github.com/nschneid/nanni/wiki/MWE-Annotation-Guidelines>

corpus documentation does not mention detailed annotation guidelines nor formal criteria, but each category contains a few examples and a brief description, along with some general annotation principles.

PolyCorp (Tutin *et al.* 2016; Tutin and Esperança-Rodier 2019) is a French corpus annotated with MWEs and NEs comprising almost 70,000 tokens from various genres. A lexicon of 5,000 MWEs, compiled from different sources, has been used to pre-identify MWEs, which were then classified as literal versus idiomatic. Expert annotators also completed the annotation with MWEs not present in the dictionary, and with NEs. The typology of MWEs builds on Mel'čuk (2012) (Section 2.2), and includes pragmatic MWEs (e.g. *you're welcome*). Although the annotation guidelines provide rough definitions of the MWE categories, they lack operational criteria for the identification task.⁷

Savary *et al.* (2017) and Ramisch *et al.* (2018) present two releases of multilingual corpora annotated for verbal MWEs in 18 (resp. 20) languages belonging to more than 5 language families in the framework of the PARSEME project. The corpora contain around 5.4M (resp. 6.1M) tokens, 62k (resp. 79k) occurrences of verbal MWEs, distributed over 5 (resp. 8) linguistic categories. A contribution of this work is the use of guidelines with precise decision flowcharts relying on linguistic tests, which have proved to be robust across languages. We summarize them in Section 6, as our work actually builds on the PARSEME annotation: we reuse the French part of the PARSEME 1.1 annotations of verbal MWEs (those made on the Sequoia corpus), and further annotate all other categories of MWEs.⁸

MWE annotation within treebanks

2.3.2

While treebanks are quite numerous, treebanks including consistent MWE annotation are rarer. Annotation guidelines for MWEs are more or less detailed depending on the project's focus. Rosén *et al.* (2015) present a survey on MWEs in treebanks. The 17 investigated treebanks have different annotation schemes and heterogeneous coverage in terms of MWE categories. Overall, one take-away message is

⁷ We thank Agnès Tutin for sending us the PolyCorp annotation guidelines.

⁸ Only a few corrections were made to the PARSEME annotation.

that better documentation of treebanks is needed, including annotation guidelines and tagsets, to help interpret MWE annotations.

We now detail some treebanks whose authors make substantial efforts to consistently annotate MWEs. The French treebank (Abeillé et al. 2003, 2019) contains about 20,000 sentences from the *Le Monde* newspaper, with MWEs annotated on top of morphological and syntactic layers. The annotation guide (Abeillé and Clément 1999–2015) lists a number of generic graphical, morphological, syntactic and semantic properties of MWEs. These are explicitly considered neither sufficient nor necessary, but should be used to evaluate whether there is sufficient evidence for the MWE status. Additionally, a typology based on the MWE’s part of speech is proposed with 8 main types (multiword nouns, pronouns, determiners, adjectives, prepositions, adverbs, conjunctions and verbs) and 10 subtypes. Some hints are given as to the choice among competing types (e.g. multiword adjectives vs. nouns vs. adverbs, etc.). The annotated verbal MWEs are limited to those which exhibit no flexibility or contain cranberry words. Formally, the annotated MWEs are almost all continuous.⁹ No evaluation of the MWE annotation quality was carried out. In the context of joint MWE identification and syntactic parsing, Candito and Constant (2014) have automatically remodeled the dependency version of the French treebank so that syntactically regular MWEs get a regular syntactic structure. MWE status is indicated using features. We have retained this principle in the MWE annotation of the Sequoia corpus (Section 9).

The Prague Dependency Treebank (Hajič et al. 2017) is a project for the Czech language started in the nineties. Several layers of annotation are defined, with MWE annotation appearing at the level of the tectogrammatical layer, which abstracts away from grammatical marking (Mikulová et al. 2006; Bejček and Straňák 2010). Tectogrammatical layers contain nodes corresponding to semantically full lexemes, potentially realized as MWEs in lower layers. The guidelines consist of examples of various MWE categories (Mikulová et al. 2006). They contain precise definitions for some MWE categories, such as verbal MWEs containing reflexive markers, or numerals, but for other

⁹ Discontinuity is allowed according to the guidelines, but among the 32 thousand annotated instances, only 59 are discontinuous (Abeillé et al. 2019).

cases, the guidelines focus on how to annotate once an MWE is identified, and do not contain operational tests nor criteria.

Universal Dependencies is an international initiative to collectively construct a highly multilingual set of syntactic-dependency treebanks using the same annotation guidelines, while leaving some space for language specificities (Nivre *et al.* 2016). For instance, version 2.5 comprises 157 treebanks and 90 languages. The annotation guidelines have a section devoted to MWEs, limited to three categories: fixed grammaticalized expressions (e.g., *in spite of*), exocentric semi-fixed expressions (e.g., *Barak Obama*) and endocentric compounds (e.g., *noun phrase*). Each category is roughly defined without operational criteria.

Named entity annotation

2.4

Named entity annotation has a long-standing tradition, notably because of the high semantic charge of NEs in texts, and thus their crucial role in semantically-oriented applications such as information extraction and sentiment analysis. The high popularity of NEs in NLP tasks was initiated by the MUC conferences (Chinchor 1998) in English, and by the benchmark for multilingual NE recognition established by the CoNLL shared tasks (Tjong Kim Sang 2002; Tjong Kim Sang and De Meulder 2003).¹⁰ This benchmark consists of datasets in Dutch, English, German and Spanish with 13,000, 35,000, 20,000 and 18,000 annotated NEs, respectively, mainly person, organization and location names; as well as some NEs of other categories, aggregated as “miscellaneous”. In these corpora, the annotation schema is rather simple: 4 main categories are used, nested NEs are not distinguished, and metonymy (e.g., person names used as names of companies) is disregarded, that is, only the effective NE categories (here: organization) are indicated. However, the 2003 CoNLL shared task edition acknowledged the interaction between syntax and NEs, in that the NE annotation is accompanied by a parallel annotation layer dedicated to chunks.

¹⁰ Available at <https://www.clips.uantwerpen.be/conll2002/ner/> and <https://www.clips.uantwerpen.be/conll2003/ner/>.

The complexity of the syntax-NE interplay lies in the fact that some NEs form a sublanguage with specific, though regular, syntactic rules. For instance, in French it is hard to identify the headword in complex person names (*Mr. Joël Bucher*) or addresses (*Jean Jaurès Str. 3*) because, differently from other languages like Greek or Polish, there is no morphological agreement hinting at a name's internal structure. Also, passages in a foreign language cannot be analysed by the grammar of the main language of a treebank (Bejček et al. 2011). Notably for these reasons, NEs are often addressed jointly with syntax in treebanks (Rosén et al. 2015). Namely, as many as 16 treebanks in 14 languages report on at least a partial coverage of NEs in their annotations. In the simplest cases, components of continuous NEs are merged into single tokens (*Alejandro_Couceiro*). If NE components are kept as separate tokens, NEs can form flat subtrees marked with uniform labels (e.g., the name relation in Universal Dependencies).¹¹ In more elaborate annotation schemas, the NE marking belongs to a different annotation layer than syntax, the NE typology includes several categories and subcategories, and nested NEs are identified (Savary et al. 2010). Finally, NEs can also be represented in the deep syntactic layer, built upon the surface syntactic layer, so that morphosyntactic variation, ellipsis and discontinuity are neutralised (Bejček et al. 2011). NEs annotated in treebanks can be further interlinked with their lexical entries (Bejček and Straňák 2010), allowing coreference markup.

A more comprehensive account of NE-annotated corpora worldwide is beyond the scope of this article.¹² Unfortunately, hardly any NE annotation guidelines are accessible online. Those few which could be accessed at the time of writing are often mainly repositories of NE categories to account for and examples to illustrate them, as well as more precise guidelines about a NE's span in text (e.g., inclusion of qualifiers and titles). We found no guidelines in which tests and decision flowcharts guide the annotator, as in our guidelines (Section 5).

Concerning French, one of the most advanced NE annotation projects was undertaken for the 1.4-million-word Quaero corpus (Grouin et al. 2011) of transcribed speech, manually annotated with

¹¹ <https://universaldependencies.org/docs/en/dep/name.html>

¹² A list of 177 such resources in 34 languages, documented with 16 attributes, can be found at <http://damien.nouvelles.net/resourcesen/corpora.html>.

a NE taxonomy of 7 categories and 32 subcategories. There, complex NEs are not only marked for nesting but also for fine-grained categories of internal components such as `name.last`, `zip-code`, `month`, etc. Also, metonymy is accounted for by primitive and effective categories (Section 5.1).¹³ While the Quaero corpus is not openly available, its biomedical spin-off corpus, inspired by the same guidelines, is distributed under an open license (Névéal *et al.* 2014). It contains more than 100,000 words and 26,409 entity annotations mapped to 5,797 unique concepts of the UMLS ontology. Another French resource, the French Treebank, was extended with about 11,000 NE annotations by Sagot *et al.* (2012). Their typology contains 7 main categories and a number of subcategories, but nested NEs were disregarded. Some of their pairs of categories correspond to a single one in our tagset. Their seven categories have the same coverage as our four coarser categories ORG, LOC, PERS, PROD. Conventions on NE spanning are very similar to ours. This resource includes an additional feature compared with our work: each mention of NE is linked to the entity database Aleda (Sagot and Stern 2012). The annotation process consisted of an automatic pre-annotation followed by a manual correction/validation by a single annotator. No quality evaluation of the resource was performed. This corpus is available for research under a specific license.

MAIN DISTINCTIONS IN PARSEME-FR TYPOLOGIES

3

Both for organizational and scientific reasons, we design our guidelines along two primary distinctions. First, we set aside verbal MWEs, which were already annotated within the multilingual PARSEME network (Section 3.1). Second, we distinguish between NEs and MWEs (Section 3.2). This results in two typologies and three categories of annotated expressions: NEs, verbal MWEs and non-verbal MWEs (Section 3.3).

¹³See the Quaero annotation guidelines at <http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf>.

3.1 *Building on the verbal MWE annotation from PARSEME*

The identification of *verbal* multiword expressions (VMWEs) has been the focus of the PARSEME shared tasks (Savary et al. 2017; Ramisch et al. 2018), initiated within the PARSEME European COST project. The PARSEME 1.1 guide for VMWEs was designed and used to produce annotations for 20 languages, including French.¹⁴ Four of the five defined categories of VMWEs are relevant for French (detailed in Section 6). We thus focused on other MWEs (non-verbal MWEs), and simply imported the existing annotations of VMWEs from PARSEME. Since members of the French spin-off project PARSEME-FR were highly involved in designing the multilingual PARSEME guide, both guides are similar in spirit.

3.2 *Distinguishing NEs from nominal MWEs*

For nominal expressions, we make a primary distinction between NEs and MWEs. A first motivation for this distinction is that, roughly speaking, most categories of NEs are inherently more productive than MWEs, and thus the latter are more suitable to be listed in a lexicon. Secondly, although both categories do share some properties that can be used in identification criteria, we found it simpler to use distinct guidelines. Moreover, we annotate both multiword and single-word NEs, since excluding the latter would have reduced the usefulness of the annotated corpus.

The NE versus MWE distinction concerns the naming convention linking an expression and the entity (or entities) it refers to. The starting distinction among nominal expressions is between a name assigned to an instantiation of a category versus a name assigned to a category (and used to refer to the category or more frequently to instances of this category):

- (A) The nominal expression e is the direct name of an entity (for instance $[Anna\ Duval]_{PERS}$),¹⁵ “direct” meaning here that the entity name is not at the same time the name of a concept which this

¹⁴<https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/>

¹⁵NEs appear in square brackets with a subscript category code (Section 5.1).

entity is an instantiation of. The name *e* may well be ambiguous (namely there can be several women named *Anna Duval*), but the key aspect is that a speaker must learn a naming convention for each entity bearing that name (Kleiber 2007). Even though a speaker knows a person *x* named *Anna Duval*, when meeting a new person *y* named that way, the speaker cannot guess her name, and has to learn the specific naming convention between *y* and the name.

- (B) The nominal expression *e* is an instantiable concept name, which can be used to refer to a concept or more often to instances of this concept (e.g., the simple noun *table* or the compound *neural network*). A naming convention does exist, but it links the name and the concept. Knowing the defining characteristics of the concept enables a speaker to use *e* to name previously unknown instances of that concept, without the need to learn any new naming convention. For instance a speaker can use the noun *table* to name a previously unseen table.

Like entity names, compositional noun phrases may unambiguously refer to entities, whether independently of the linguistic context (e.g., *the first British female prime minister*) or thanks to the context (e.g., *the woman* used for a specific woman, disambiguated in context). However, as opposed to entity names, the reference of compositional noun phrases is momentary, not intended to last (Kleiber 2007).

The distinction between entity direct names and instantiable concept names is reminiscent of the proper noun versus common noun distinction, but the latter proves not so easy to draw. Of course, lexical items that are exclusively devoted to directly naming entities (e.g., the first and last names for people) are easily classified as proper nouns (sometimes called *pure* proper nouns). This is why Ehrmann (2008) roughly defines proper nouns as “the designation of a precise entity via a description whose meaning plays a minor role with respect to the denomination of the referent, which operates directly”.¹⁶ However, abundant literature shows that the proper vs. common noun distinction is difficult to characterize in linguistic terms (Kleiber 2001, 2007; Ehrmann 2008). Within direct names of entities, we rather distinguish:

¹⁶Translated from French (Ehrmann 2008, p. 172).

- (A₁) names made of lexical items dedicated to naming specific entities (pure proper nouns), such as *[Italy]_{LOC}* and *[Anna Duval]_{PERS}*;
- (A₂) names that are semantically compositional, either totally (such as the *[International League against Racism and Anti-Semitism]_{ORG}*) or partially (such as *[massif central]_{LOC}* ‘central massif’, referring to a specific massif at the center of France, or *[mer de glace]_{LOC}* ‘sea of ice’ for a specific glacier in the Alps); the important feature, though, is that these are names of specific entities for which a direct naming convention must be learnt;
- (A₃) names which designate unique abstract entities, such as abstract simple nouns (*taxidermy*) or abstract MWEs (*Euclidean geometry*, *natural language processing*): because of the unicity of the entity that can be called that way, they too can be viewed as entity names, for which the speakers have to learn the naming convention at the level of the entity.

However, cases (A₃) are traditionally not viewed as proper nouns. Kleiber (1996) argues that pure proper nouns are meant to name a particular entity within a well-identified semantic class (e.g., a person), whereas for (A₃) cases, the relevant hypernym is not obvious. We have chosen to follow this tradition, considering cases (A₁) and (A₂) as proper nouns, and (A₃) as common nouns. In short, we distinguish:

- **NEs:** We tag cases (A₁) and (A₂) as *named entities* and associate them with a semantic category. Although the term is confusing (one should speak of an entity name, not a named entity) we use it for entity names, as it is usual in the NLP community. We annotate these as NEs using dedicated guidelines (Section 5).
- **MWEs:** We tag as *multiword expressions* cases (B) and (A₃), provided they are composed of more than one component.

Finally, there are also names referring to unique concrete entities such as the sun or the moon (often called “unica”), whose status is widely debated. We have chosen to tag these as NEs (e.g., *I can see you thanks to the [moon]_{LOC}*), unless when it is clear they refer to a concept instance (e.g., *Many planets have a moon*).

The MWE vs. NE dichotomy is particularly challenging due to at least three facts. Firstly, MWEs can contain NEs, as in *maladie de*

[Paget]_{PERS} ‘Paget’s disease’ and vice versa **[Association nationale des anciens combattants de la Résistance]**_{ORG} ‘Association of the Old Fighters of the Resistance’⇒‘Resistance Veteran Association’.¹⁷ Secondly, due to ellipsis, an NE can boil down to those components which form an MWE, e.g., **[Anciens combattant]**_{ORG} ‘Old fighters’⇒‘Veterans’ can either refer to a class of people or be a shortcut for the full organization name. Our guidelines, however, exclude annotating a sequence both as NE and MWE (here, only the NE annotation applies). Thirdly, as pointed out above, many NEs have a descriptive basis, e.g., **[Cour d’appel de [Paris]**_{LOC}**]**_{ORG} ‘Court of Appeal of Paris’, and their status as NEs stems from the naming convention, possibly specific to a particular domain of expertise (e.g., law) not familiar to the annotators. Given these challenges, we formalized dedicated decision flowcharts, discussed in Sections 4.2, 5.3 and 6, so as to maximise the reproducibility of the process.

PARSEME-FR typologies

3.3

The typologies resulting from the distinctions explained above and used in our annotation are depicted in Figure 1. NEs are split into 5 categories, and MWEs divide into non-verbal MWEs – subdivided into syntactically regular and irregular (Section 9) – and VMWEs, with 4 relevant categories and 2 subcategories inherited from PARSEME.

Comparing these typologies to the ones described in Section 2.2, several facts are worth noting. Firstly, like Sag *et al.* (2002) and Tutin and Esperança-Rodier (2019), we model and annotate MWEs and NEs in the same framework. However, unlike these two previous works, we distinguish named entities and MWEs. More precisely we make a semantic difference concerning the level at which the naming convention operates (cf. Section 3.2), and hence we consider the MWE typology as disjoint from the NE typology, the latter including both single- and multi-word NEs.

Secondly, our typologies are heterogeneous, as we define NE and MWE subtypes using different criteria. The typology of NEs is based

¹⁷In examples, components of MWEs are shown in bold. Idiomatic translations of MWEs in inline examples, when required, are preceded by an arrow ⇒.

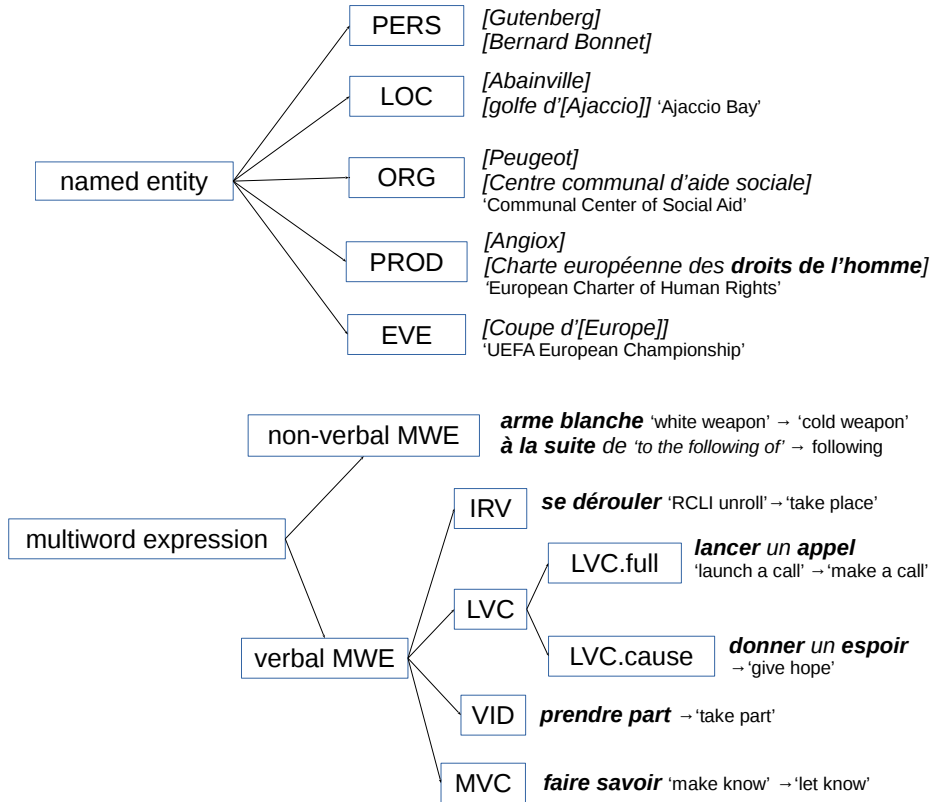


Figure 1: Named entity and multiword expression typologies used in the PARSEME-FR corpus

on the semantic types of the named objects and ignores the linguistic properties of the names themselves. Conversely, the MWE typology is largely driven by the syntactic structure of the annotated expressions. Also, while verbal MWEs are further divided into finer subtypes, non-verbal MWEs are not. This situation results from a mixture of historical and linguistic factors. NE annotation has a long-standing tradition and opposing it in such fundamental aspects as typology design principles might jeopardise the utility of the corpus. In particular, annotating single-word NEs seemed valuable from an applicative perspective. The PARSEME typology and guidelines are exclusively dedicated to verbal MWEs but have the advantage of being validated in a multilingual framework. Their elaboration is justified by the fact

that VMWEs show a relatively high degree of syntactic flexibility and discontinuity. Thus, to make the guidelines operational, the syntactic tests included therein must be structure-specific. For non-verbal MWEs, such structure-specific guidelines proved unnecessary in our experience. What is more, when defining the syntactic categories for non-verbal MWEs, we would have to face hard challenges,¹⁸ not central to our interests. Note however that considerable effort was dedicated to part-of-speech annotation for syntactically irregular MWEs (cf. Section 9).

Thirdly, our NE typology is coarser than in some previous efforts dedicated to NEs alone, notably in the French corpus by Gravier *et al.* (2012) with 7 categories and 32 subcategories. Also, while other NE-dedicated efforts cover temporal expressions (e.g., dates) and measures (e.g., amounts of money), we exclude them from our annotation scope, because we believe that, while they stem from specific grammatical subsystems, their semantics remain compositional and require no entity-specific naming convention (Section 3.2).

Fourthly, our annotation scope does not cover collocations, which we define as word combinations whose idiosyncrasy is of statistical nature only (e.g., *drastically drop*). However, what other projects call collocations is partly included in our scope. For instance, our light-verb constructions cover a subset of Mel'čuk's collocations, namely those concerned by the lexical function called *Oper*.

Fifthly, the number of annotated NEs and MWEs (Section 10), exceeds 6,500 corpus occurrences, roughly balanced between NEs and MWEs, which is comparable to the work of Schneider *et al.* (2014), who however only use 2 main categories.

Finally, and most importantly, our typologies are endorsed by extensive annotation guidelines based on decision flowcharts over linguistic tests, which are meant to guide the annotator – in a relatively deterministic and reproducible way – to both identify and categorize candidate MWEs/NEs into one of the proposed categories. In particular, we largely cover the challenge of distinguishing between NEs and MWEs themselves – in terms of operational definitions, even though

¹⁸For instance, preposition-noun patterns, as in *à raison de* 'in reason of' ⇒ 'at a rate of', are notoriously hard to categorise into adjectival, adverbial or prepositional phrases.

both categories of expressions share properties. To the best of our knowledge, this constitutes an unprecedented outcome.

4 GENERAL ANNOTATION GUIDELINES

Our annotation guidelines start with a description of some formal constraints (Section 4.1) and a top decision flowchart (Section 4.2).

4.1 *Formal constraints and format*

While annotating MWEs and NEs, we face most of the annotation challenges pointed at by Mathet *et al.* (2015) and Savary *et al.* (2018):

- unitizing, that is, identifying the boundaries of the NE or MWE, which is often challenging, in particular for NEs;
- categorising (for NEs);
- free overlap, in particular in coordinated MWEs *il peut plaider_{1,2} coupable₁ ou non₂ coupable₂* ‘he can plead guilty or non guilty’.
- nesting, as in *Il a fait₁ un véritable faux pas_{1,2}* ‘he made a true false step’ \Rightarrow ‘He really made a faux pas’, which contains a light-verb construction whose predicative noun is itself a MWE.
- discontinuities (as in the previous examples).

The sole formal constraint we have put on the annotation is that we only consider MWEs that are syntactically connected, that is, whose components form a connected dependency subtree in the syntactic representation.¹⁹ A counter-example is *ce NOUN-là* ‘this NOUN-here’ \Rightarrow ‘this NOUN’.²⁰ The two potential components *ce* and *-là* syntactically depend on the noun, which is an open slot and cannot be part of the MWE.

¹⁹ More precisely, a *canonical form* of the MWE needs to form a connected dependency subtree. A canonical form of a MWE is one of its least marked syntactic forms preserving the idiomatic meaning. This mainly affects VMWEs. Note that the canonical form of a MWE is not necessarily the most frequent one.

²⁰ We use part-of-speech tags from the Universal Dependencies project.

Apart from this restriction, in a given sentence, any set of tokens can form a MWE or NE, and a given token can belong to several MWEs or NEs.

In practice, the annotations of MWEs and NEs are provided as the 11th column added to a CoNLL-U file²¹ containing morphological and syntactic annotations.²² MWEs/NEs are annotated using integer identifiers, which are sentence-specific. Additional information is provided on the first token of an MWE/NE: (i) the part of speech of the MWE, unless the MWE is considered syntactically regular (see below Section 9); (ii) the MWE versus NE category, plus the subcategory of NE or of verbal MWEs, e.g., NE-PERS or MWE-LVC; and (iii) for non-verbal MWEs: one matching sufficient criterion.

Top decision flowchart

4.2

As discussed in Section 3, the three main categories of expressions in our typologies are NEs, VMWEs and non-verbal MWEs, each of which is covered by separate annotation guidelines. Figure 2 shows the top decision flowchart²³ which guides the annotator to the appropriate guidelines.

The initial step (CAND) of identifying a potential expression to annotate is largely based on the annotator's intuition, which is further confirmed or contradicted by more rigorous guidelines. In this step, a candidate *c* can be composed of one or more lexemes since single-word NEs are also annotated.²⁴

²¹<https://universaldependencies.org/format.html>

²²The precise description of the format is available at <https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Corpus-format-description>

²³https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Guide-annotation-PARSEME_FR-chapeau#top-decision-tree

²⁴Lexemes are only roughly approximated by tokens, depending on the corpus tokenization. We use the original tokenization of the corpus, but consider certain tokens as multiword if they contain non-alphanumeric characters, annotating them as MWEs when the guidelines apply, e.g., *peut-être* 'may-be'⇒'maybe'.

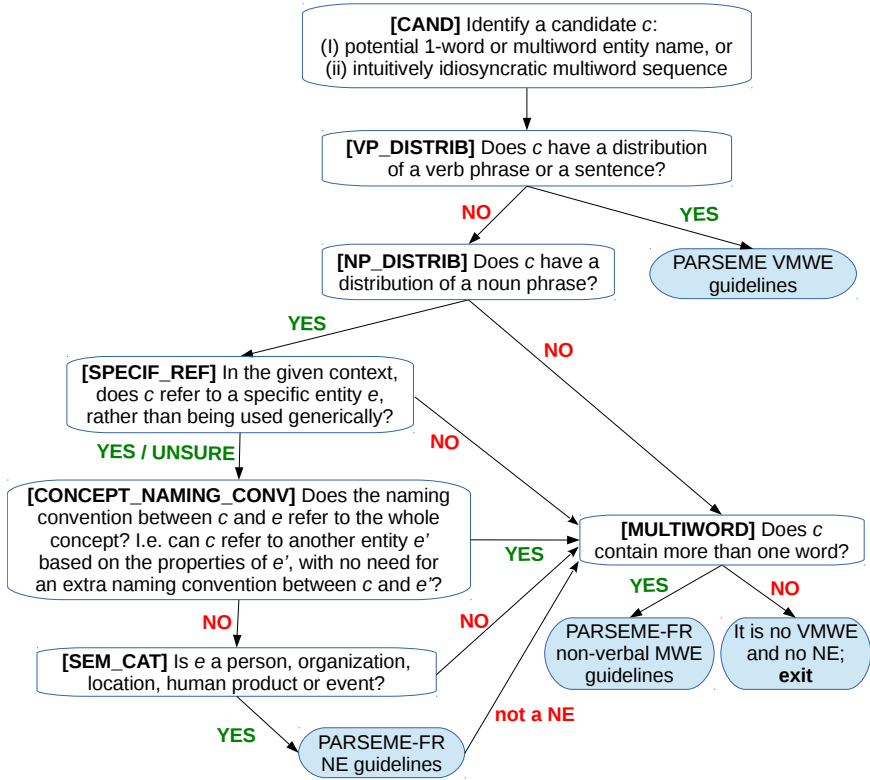


Figure 2: Top decision flowchart of the annotation guidelines

The next step (VP_DISTRIB) redirects to the PARSEME VMWE guidelines if c has a distribution of a verbal phrase or a sentence, e.g. *il vide son sac* ‘he empties his bag’ \Rightarrow ‘he gets it off his chest’.²⁵

If c is neither verbal nor nominal (NP_DISTRIB), e.g., *à l’issue de* ‘at the outcome of’ \Rightarrow ‘after’, it is tested against our non-verbal MWE guidelines, provided that it is composed of two or more lexemes, and discarded otherwise.²⁶

If c is nominal, it can (in the given context) either be used generically, as in (1), or refer to a specific entity e (SPECIF_REF), as in (2).

²⁵ <https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/>

²⁶ <https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/-/wikis/Criteres>

- (1) Le **conseil régional** est l'assemblée délibérante d'une région.
'The general council is the deliberating assembly of a region.'
- (2) Le **conseil régional** a délibéré hier soir.
'The general council deliberated last night.'

In the former case, *c* cannot be a NE but, if multiword, it might be a non-verbal MWE. In the latter case (or if the test is hard to apply), it is necessary to determine the naming convention which links *c* to its referent *e*. If this convention covers the whole concept (CONCEPT_NAMING_CONV), as in (2), then *c* can (in other contexts) refer to another referent *e'* on the basis of the properties of *e'*. In this case, if *c* is multiword, it might be a non-verbal MWE.

Conversely, the naming convention may cover only the link between *c* and *e*, rather than a whole concept. In this case, one of the two possibilities arises: (i) *c* can refer to another referent *e'* only if a new naming convention is established, as in [*Anna Duval*]_{PERS}, or (ii) *e* is, by nature, unique, so there can be no other *e'* which *c* can refer to, as in *physique quantique* 'quantum physics' or in [*Journal officiel de la République française*]_{ORG}_{PROD} 'Official Journal of the French Republic'. In any of these two cases *c* might be an NE. Thus, if *e* belongs to one of the pre-selected semantic categories (person, organization, location, human product or event), then *c* is tested against the PARSEME-FR NE guidelines. If their outcome is negative and if *c* is multiword, it might still be a non-verbal MWE.

The SPECIF_REF and CONCEPT_NAMING_CONV tests are meant to distinguish cases (A) and (B) from Section 3.2. The distinction between cases (A₁) and (A₂) on the one hand, and (A₃) on the other hand, is implemented by the SEM_CAT test and the PARSEME-FR NE guidelines.

GUIDELINES FOR NAMED ENTITIES

5

This section describes the typology (Section 5.1), principles (Section 5.2) and tests (Section 5.3) used for the annotation of NEs.

5.1

Named entity categories

The scope of the NE annotation covers the following categories:

- persons (PERS), e.g., [*Gutenberg*]_{PERS}, [*Bernard Bonnet*]_{PERS};
- locations (LOC), e.g., [*Abainville*]_{LOC} ‘a French city’, [*golfe d’Ajaccio*]_{LOC} ‘Ajaccio Bay’;
- organizations and human collectives (ORG), e.g., [*Comité départemental d’action touristique*]_{ORG} ‘Departement Committee of Tourism’;
- products, including titles of works and documents (PROD), e.g., [*Angiox*]_{PROD}, [*Charte européenne des droits de l’homme*]_{PROD} ‘European Charter of Human Rights’, [*Libération*]_{PROD} ‘a newspaper’;
- named events (EVE), e.g., [*L’affaire [Dumas]*]_{PERS} ‘Dumasgate’, [*Coupe d’Europe*]_{LOC} ‘UEFA European Championship’.

Dates, amounts, and numerical expressions, commonly covered by the NE term in the NLP literature (e.g., in the work of Chinchor (1997) followed by Tjong Kim Sang and De Meulder (2003)) are not included in this scope, since they do not name a specific entity in the discourse world.

A pervasive feature of NEs is that they occur as metonyms, in which case a change of NE category frequently occurs. Since metonymy is one of the hardest challenges in NE recognition (Markert and Nissim 2007), we account for it in the annotation schema. For metonymic uses of NEs, we mark both the effective (called *final*) and the primitive NE category. For instance, in *chauffeur-routier chez [Caillaud]*_{PERS}^{ORG} ‘truck driver from Caillaud’, the last token *Caillaud* is originally the name of a person, further assigned to a company. Thus, the primitive and the final categories are PERS and ORG, respectively.²⁷ In some cases it is hard to decide which of the two considered types is primary or final. For instance, we may hesitate between considering a journal name as primary and its editorial office as final, or vice versa. In such controversial cases, we follow the default priority order LOC < PERS < ORG < PROD (where < means less final, more primitive). For instance, in *informations publiées dans*

²⁷ We use a superscript to indicate the primitive category.

[*Le Canard enchaîné*]_{PROD}^{ORG} ‘information published in The Chained Duck (a newspaper)’ we indicate both the primary and the final type. Conversely, in *accusation portée par [Le Canard enchaîné]*_{ORG}^{ORG} ‘accusation brought by The Chained Duck’ only the final type appears (i.e. there is no metonymy).²⁸

A NE can undergo a series of metonymies, in which case we only mark as primitive the category which directly precedes the final category in this series. For instance, in [*Reuters*]_{PROD}^{ORG} the surname (PERS) of the founder *Paul Reuter* of the press agency (ORG) further became the name of the released informational content (PROD). Here, only the last two categories are annotated as primary and final, respectively.

Note also that metonymy can invalidate the NE status in some cases. Notably, trade marks used metonymically (to refer to products themselves), e.g., *BMW* in [*Anna*]_{PERS} *a acheté une BMW* ‘Anna has bought a BMW’, are not annotated as NEs.²⁹ Here, the naming convention (addressed by the CONCEPT_NAMING_CONV test in Section 4.2) between a particular car and the *BMW* name need not be re-established, but stems from the car’s properties instead.

Nested and overlapping named entities

5.2

NEs frequently exhibit nesting, with or without intervening MWEs. We annotate all these nested instances, as in [*Cour d’appel de Paris*]_{LOC}^{ORG} ‘Court of Appeal of Paris’, which implies that some tokens belong to several annotated entities. Note that in people’s names like [*Jean-Paul Alègre*]_{PERS} the given names and surnames are no autonomous nested NEs but rather ellipses of the full names, or *components* (Grouin *et al.* 2011), therefore they are not to be annotated separately.

²⁸Note that primitive types are marked only in case of a clear metonymic relation between the referenced objects (part/whole, container/contents, cause/effect, artist/work, location/inhabitants, location/institution, etc.). Other cases of polysemy are not relevant, e.g. when a place is named after a person (*Washington*_{LOC}) or a god (*Mars*_{LOC}).

²⁹An alternative approach would have been to annotate *BMW* as a NE with the primitive category (PROD) only, but we favor overall coherence instead.

Another case of overlapping annotations stems from coordinations, as in *les traités*_{PROD₁,PROD₂} *de*_{PROD₁} *Rome*_{PROD₁,LOC₁} *et de*_{PROD₂} *Paris*_{PROD₂,LOC₂} ‘treaties of Rome and of Paris’, where some components of the annotated entities are shared (here: *traités* ‘treaties’).³⁰

5.3

Linguistic tests and decision flowchart

The topmost decision process in the PARSEME-FR guidelines (Section 4.2) branches to the NE guidelines when the candidate expression refers to a specific discourse entity in context and there might be a naming convention linking this expression with this particular entity. In order to confirm an intuition that the annotator may have about the candidate at hand, the NE guidelines are organized as a decision flowchart, so as to maximize the reproducibility of the annotator’s decisions.³¹

The two main challenges to be faced here are: (i) identifying the naming convention concerning the NE candidate at hand, and (ii) determining the textual span of the candidate. Stage (i) is handled by the following linguistic tests:³²

- **OBVIOUSPROPER**: Is the candidate sequence obviously a proper name, that is, is the annotator confident about the existence of the naming convention concerning the sequence?
- **RELEVUPPER**: Is the candidate sequence, or its variant in the same text, spelled with an initial uppercase letter to signal a proper name, rather than for other (e.g., honorific) reasons?
- **ACRON**: Does the candidate sequence have an acronym in the given text?
- **WEBPAGE**: Is there an official web page or Wikipedia page titled by the candidate sequence?

³⁰ Discontinuous NEs are marked by subscript identifiers on each component.

³¹ <https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/ne-decision-tree>

³² These tests are not applied sequentially but included within the decision flowchart mentioned above, omitted here for the sake of concision.

Stage (ii) is particularly challenging in French, because in multi-word names of organizations only the initial of the first component is usually capitalised, as in *Association paroissiale d'éducation populaire* 'Parish Association of Popular Education'. Additionally, the attachment of prepositional phrases (PPs) to NEs is notoriously hard, in particular for location PPs. Therefore, stage (ii) also relies on the available external sources via the three last tests above. Namely, if ACRON or WEBPAGE apply, the span is usually easily determined by the acronym or the title of the relevant webpage. Two additional tests dedicated to the NE span are used within the decision flowchart:

- MINSPAN: Does the candidate sequence *c* have the minimal span, that is, is it true that a shorter span than *c* no longer refers to the same entity? For instance, the test is passed for *[la Rochelle]_{LOC}* (since the determiner cannot be omitted).
- SPANPERCAT: If the preceding tests were not sufficient to determine the inclusion of the classifier, it is systematically excluded in names of persons (*colonel [Pétain]_{PERS}*), products, events, regions, departments, cities (*la ville de [Loudun]_{LOC}* 'the city of Loudun'), and some organizations (*société [Cedel]_{ORG}* 'Cedel company'). In other cases, the classifier is systematically included (*[école Notre-Dame]_{LOC}* 'Our Lady's School', *[ministère français des Affaires étrangères]_{ORG}* 'French Ministry of Foreign Affairs'). Although somewhat arbitrary, this list of cases ensures coherence for some notoriously difficult cases.

GUIDELINES FOR VERBAL MWEs

6

The annotation of verbal MWEs in the PARSEME-FR corpus is transferred from the multilingual PARSEME corpus annotated for VMWEs (Savary *et al.* 2018; Ramisch *et al.* 2018), and its French subcorpus was described in detail by Candito *et al.* (2017). Version 1.1 covers 20 languages, including French.³³ The guidelines are organized as a

³³<http://hdl.handle.net/11372/LRT-2842>

generic flowchart, based on linguistic tests, which redirect to category-specific flowcharts.³⁴ Six major categories are defined, four of which are relevant to French.

- *Inherently reflexive verbs* (IRV) are combinations of a verb *v* and a reflexive clitic *r*, such that one of the non-compositionality conditions holds: (i) *v* never occurs without *r*, like in (3); (ii) *r* distinctly changes the meaning of *v*, like in (4); (iii) *r* changes the subcategorization frame of *v*, like in (5) as opposed to (6).

(3) Je **me souviens** de ce livre.

I self remember of this book

‘I remember this book.’

(4) Une seconde opération **se déroulait** en parallèle.

a second operation self unrolled in parallel

‘Another operation was taking place at the same time.’

(5) Je **m’occupe** du dessert.

I self occupy of-the dessert

‘I take in charge the dessert.’

(6) J’occupe les enfants avec un jeu.

I occupy the kids with a game

‘I keep the children busy with a game.’

- *Light-verb constructions* (LVCs) are verb-noun combinations in which the verb is semantically void or bleached, and the noun is a predicate expressing an event or a state. Two subcategories are defined: *LVC.full* are those LVCs in which the subject of the verb is a semantic argument of the noun, as in (7); *LVC.cause* are those in which the subject of the verb is the cause of the noun (but is not its semantic argument), as in (8).

(7) Nous devons **lancer un appel** à la raison.

we must launch a call to the reason

‘We must make a call to reason.’

³⁴<http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/>

- (8) Il **donne espoir** aux soldats.
he gives hope to soldiers
'He gives hope to soldiers.'

- *Verbal idioms* (VIDs) are verb phrases of various syntactic structures which contain cranberry words or exhibit lexical, morphological or syntactic inflexibility, as in (9).

- (9) petit resto qui **ne paye pas de mine**
small restaurant which NEG pays NEG DET face
'small restaurant which is not much to look at'

- *Multi-verb constructions* (MVCs), rare in French, consist of a sequence of two verbs, so that replacing one verb by a verb from the same broad semantic class leads to ungrammaticality or to an unexpected change in meaning, as in (10).

- (10) Il n'avait jamais **entendu parler de** ça.
he NEG'had never heard talk about this
'He had never heard of this before.'

GUIDELINES FOR NON-VERBAL MWES

7

Below, we justify the use of sufficient criteria (Section 7.1), discuss annotation span (Section 7.2), and present the criteria (Section 7.3).

General principles: sufficient criteria

7.1

A specific decision flowchart³⁵ indicates whether a candidate *c* (already identified as not being a NE) is an MWE or not. The main characteristic of these guidelines is that, unless stated otherwise, *each individual criterion is sufficient to tag the candidate as an MWE*. This is intended as a solution to the well-known difficulty to make binary decisions within the continuous scale of idiomaticity. It is reminiscent of

³⁵<https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Criteres>

how the lexicon-grammar is organized using dozens of binary properties Gross (1994). The alternative solution, used e.g., for MWEs in the French Treebank, is to ask annotators to judge whether there are enough satisfied criteria in order to tag a sequence as MWE (Abeillé and Clément 1999–2015).³⁶ The number and the relative weight of the criteria being difficult to assess, we thus prefer to consider sufficient criteria only. The annotated MWEs will satisfy a varying number of criteria, thus we obtain an MWE lexicon with a varying degree of idiomaticity.

The various criteria are defined using precise linguistic tests, designed to formalize lexical, morphological, syntactic or semantic idiosyncrasy (the former being generally a clue for the last). A test generally consists of studying how a modification of *c* (such as replacing, adding or removing one component) impacts its acceptability and its interpretation. The considered modifications are only those allowed for non-MWE sequences, within the regular grammar of the language. The test succeeds if the modification leads to unacceptability: for example, in (11), the adverb *bien* ‘well’ can normally be modified by the intensifier *très* ‘very’, but this leads to unacceptability in the context of the MWE *bien que* ‘well that’⇒‘even though’. The test also succeeds if the result after modification remains acceptable, but exhibits an unexpected meaning shift given the applied modification (henceforth noted #). For instance, in the MWE *carte bleue* ‘card blue’⇒‘credit card’, substituting the color adjective by another color is acceptable, but the meaning change is not the expected change of color meaning.

- (11) Je continue (*très) **bien que** j’ ai peur.
I continue (very) well that I have fear
‘I go on (*very) even though I am afraid.’

Meaning shift is not a binary property, but rather a fuzzy value in a continuum.³⁷ A transformation applied to any phrase may yield a result which ranges from completely expected to totally surprising, with many possible interpretations in between. Ideally, we would like

³⁶The guidelines mention “a beam of criteria” (“un faisceau de critères”).

³⁷One can argue that the same is true for acceptability, although the predictability of a meaning shift is arguably more subtle to assess than a sequence’s acceptability for an average speaker.

to quantify meaning shift, e.g. as the branch distance in Wordnet, or the embeddings' cosine similarity. This would allow us to establish a numerical threshold beyond which meaning shift is considered unexpected, making annotation more reproducible. In practice, though, this is not feasible because our tests operate on whole multiword phrases, whose representation is not straightforward. We resort to comparing the same transformation to other phrases which are clearly not MWEs, and assessing whether the transformation applied to the candidate follows the same pattern, in which case it should not be annotated as an MWE, or if the meaning change is indeed unexpected with respect to similar non-MWE phrases.

Span of MWEs

7.2

When a candidate sequence passes at least one MWE test, it remains to decide which elements are actually part of the MWE (Savary *et al.* 2018). These elements do not vary lexically, that is, their lemma cannot vary (morphological variation is possible). For instance, in the sequence *en termes économiques/pratiques/démographiques* ('in economic/practical/demographic terms'), we consider *en termes* as forming an MWE, with an open slot.

Selected prepositions and complementizers introducing open slots

7.2.1

In some cases an MWE selects an argument (mandatory or not) that is not itself frozen, but is introduced by a frozen preposition or complementizer that functions as a grammatical marker. Although the marker is frozen, we have chosen not to include it in the MWE. For instance in example (12), we annotate *en* and *dépit* as an MWE, which takes a mandatory prepositional phrase with the preposition *de*, not included in the MWE. This treatment derives from the general treatment of grammatical markers: we do not consider that a verb plus the preposition it subcategorizes for forms an MWE (e.g., we do not annotate any MWE in *Je compte sur toi* 'I am counting on you', even though the preposition is frozen). Our choice is to privilege a consistent treatment of selected prepositions and complementizers at the expense of excluding some mandatory elements from the MWEs' annotation span.

- (12) Il a continué **en dépit** de nos appels.
 he has continued in bitterness of our calls
 ‘He continued in spite of our calls.’

The rule for excluding final grammatical markers has an exception, though. For a sequence containing just one component plus a selected preposition, we annotate it as MWE if it satisfies other criteria than the fixedness of the preposition. This is the case, for instance, for *faute de* ‘fault of’: it functions as a sentence modifier (13), which is normally not the case for a non-temporal noun such as *faute* ‘fault’.

- (13) **Faute d’** accord, la proposition de loi est rejetée.
 fault of agreement, the proposition of law is rejected
 ‘Since no agreement is reached, the proposed law is rejected.’

For selected complementizers, we generally follow the same rule as for selected prepositions. In particular, prepositions introducing a clause starting by *que* ‘that’ do not form an MWE with the complementizer. Indeed, in this particular case, the finite clause introduced by the complementizer generally alternates with an infinitival clause introduced by *de*, and is generally optional, as in (14). This fact provides an additional justification for not including the complementizer, and thus not annotating the combination as an MWE.

- (14) Il part avant (\emptyset | la fin | de finir | que tu finisses).
 he leaves before (\emptyset | the end | of to-end | that you end)
 ‘He leaves before (\emptyset | the end | finishing | you finish).’

As an exception, we consider certain sequences of the form ADV + *que*, as irregular and tag them MWEs (Section 12).

7.2.2

Determiners

The inclusion of a determiner in the annotation span depends on its frozen status. By default, if the determiner is totally frozen, or can vary only in gender, number, or person of the possessor, then it should be included. For instance, in *fruit de la passion* ‘fruit of the passion’ \Rightarrow ‘passion fruit’, the determiner does not accept any variation *#fruit de (cette | une | ma) passion* ‘fruit of (this | a | my) passion’.

However, deciding whether a determiner is frozen is not straightforward because we must deal with a large number of special cases. Therefore, a dedicated decision flowchart and detailed instructions, also covering the special case of “zero” determiners, are presented in the identification criteria named DET and ZERO in Section 7.3.

Identification criteria

7.3

The criteria to determine whether *c* is a non-verbal MWE are summarized as follows:

1. Semantic criteria

- [ID] the syntactic head of *c* is not its “hypernym”
- [PRED] no predication relation between head and modifier

2. Lexical fixedness criteria

- [CRAN] *c* contains a cranberry word
- [LEX] no replacement of a content word by a similar word
- [DET] the determiner of a noun is totally fixed
- [ZERO] possible empty determiner, while usually required

3. Morphosyntactic fixedness criteria

- [MORPHO] no modification of the morphological features
- [IRREG] irregular morphosyntactic structure
- [SYNT] impossibility of syntactic variation for some patterns
- [INSERT] no insertion of modifiers, while usually possible

The description of each criterion is provided in Appendix.

ANNOTATION PROCESS AND QUALITY

8

We now detail our source corpus (Section 8.1), annotation process (Section 8.2) and the quality of the MWE/NE annotations (Section 8.3).

8.1

Source corpus

We chose to annotate the Sequoia corpus (Candito and Seddah 2012), which is a freely available corpus containing 3,099 sentences, initially annotated for morphology and syntax. Other kinds of annotations were subsequently added (e.g., deep syntax and semantic frames, coarse semantic categories for nouns), thus making the overall corpus richly annotated.

The corpus was first created to perform domain adaptation experiments, hence it comprises sentences originating from four different sources: a regional newspaper (*L'Est Républicain*, narrative historical pages from the French wikipedia, Europarl transcriptions, and two medical reports from the European Medicine Agency). In the original morphosyntactic annotations, only functional MWEs had been annotated. We ignored these annotations in our first annotation phase, and used them afterwards to spot potential errors (Section 8.2).

8.2

Annotation process

Our annotation process classically comprises a pilot phase to test and improve the guidelines, a double annotation plus adjudication phase, and a further phase of coherence checking.

We chose not to use any pre-annotating tools, which are known to introduce task-dependent biases (e.g. Fort and Sagot 2010 for POS tagging). Indeed, although such tools speed up annotation and uniformize simple repetitive annotations, the negative effect is that annotators will tend to reproduce noise and silence induced by the tool (Savary *et al.* 2018). Moreover, since our main objective was to operationalize and test MWE identification criteria, we did not want to rely on pre-annotating tools, necessarily based on pre-existing MWE resources. This has obviously prevented us from annotating a large corpus.

After a first rough version of the guidelines, we performed a pilot annotation on a fraction of the Sequoia corpus, corresponding to two French wikipedia pages (containing about 2,000 tokens and 93 sentences). Four annotators (among the authors of this article) annotated this fraction, and collectively adjudicated it, gathering feedback to complete and amend the guidelines.

We then performed a double annotation and adjudication of the rest of the Sequoia corpus.³⁸ We used the FLAT tool³⁹ for annotation, with a predefined set of categories (van Gompel and Reynaert 2013). For NEs, annotators had to choose the semantic category, and annotate both the primary and final categories in case of metonymy. For non-verbal MWEs, annotators had to provide one of the sufficient criteria (listed in Section 7).

For the adjudication, we used a specific in-house tool, which showed the two parallel annotations side by side and allowed for the resolution of conflicts, namely when (a) one annotation did not have a paired counterpart; (b) it did but the sets of tokens were not the same; or (c) all tokens coincided but the assigned category differed.

After adjudication, the corpus also underwent a consistency check using a tool from the PARSEME shared-task, which extracts all annotations and clusters them. More precisely, each cluster contains the annotations of a given entity and of other entities with similar verbs or nouns, as well as non-annotated co-occurrences of words resembling this annotation. Two experts manually checked all clusters, minimizing inconsistencies and reducing both noise and silence in the corpus.

The last systematic check consisted in comparing the pre-existing annotations of functional MWEs and proper names in Sequoia (Section 8.1) against our annotations. In the end, syntactic annotation was modified to comply with our MWE annotations (Section 9).

Corpus quality

8.3

To evaluate the quality of the annotations, a common practice is to calculate the inter-annotator agreement. A popular metric to this end is the kappa score (Cohen 1960), defined for categorization tasks and evaluating the observed agreement with respect to what could be expected by pure chance. The adaptation of the kappa score to our annotations is not straightforward. One naive solution is to work at the level of tokens, considering binary decisions as to whether the

³⁸Two of the annotators are not native speakers of French, although living in France for many years. Adjudication was performed by native speakers only.

³⁹<http://github.com/proycon/flat>, <http://flat.science.ru.nl>

token belongs to an MWE/NE or not. Yet, in such a setting, the resulting agreements (both observed and by chance) would be biased because tokens not belonging to MWEs or NEs are much more frequent. Bejček and Straňák (2010) proposed an adaptation of Cohen’s kappa to measure the agreement for MWEs and NEs in the Prague Dependency Treebank. They consider annotation agreement over each syntactic tree node (representing a set of tokens in the surface sentence), and provide a complex system of weights for the various cases of (dis)agreement.⁴⁰ These weights are used to compute both the observed and the chance agreement, and hence a kappa score. Another metric is the gamma score (Mathet *et al.* 2015), suitable for unitizing tasks, that is, in which annotators have to identify by themselves which elements to annotate. Gamma is not defined, though, when the units to identify are potentially discontinuous, and when a token can belong to several units.

Hence, instead of a chance-corrected measure, we use the plain F-score between two annotations to evaluate the annotation quality, for two main reasons: firstly, because there are almost no formal constraints for MWE/NE annotation, intuitively the chance agreement is very low. This is indeed confirmed by the chance agreement of 0.046 obtained by Bejček and Straňák (2010). Secondly, adapting the kappa score to the MWE/NE identification task requires some arbitrary choices (such as the weights in Bejček and Straňák 2010), leading to a measure that we find difficult to interpret.

Table 1 shows the F-scores between the two annotations before adjudication, computed at the level of full MWEs/NEs, on the entire corpus except for 2,000 tokens used in the pilot annotation, that is, about 56,500 tokens.⁴¹ We consider both *exact* and *partial* matches: in the former case, agreement means that exactly the same set of tokens is annotated by both annotators, ignoring the category. In the latter, two

⁴⁰ For instance, agreeing for tagging a node as part of a NE or MWE, but disagreeing on the exact NE or MWE, counts as one fourth of the full agreement.

⁴¹ This corresponds to the uncategorized MWE-based metric used for MWE identification. Token-based agreement was not assessed. We ignore VMWEs since they are copied from the PARSEME project, whose original agreement was 0.766 (Ramisch *et al.* 2018) before consistency checks, and which we only marginally modified.

	F-score between 2 annotation sets	
	Exact match	Partial match
Non-verbal MWEs	55.3	58.6
↳Regional news	62.6	65.9
↳Europarl	60.1	64.6
↳French Wikipedia	61.9	68.4
↳Medical reports	41.7	42.2
Named entities	84.0	85.7
↳Regional news	85.5	87.2
↳Europarl	84.1	84.6
↳French Wikipedia	85.7	88.1
↳Medical reports	56.4	59.8
Both	71.1	73.7

Table 1:
Inter-annotator agreement;
F-scores
between the two sets
of annotated NEs
and non-verbal MWEs,
before adjudication,
in exact or partial match

annotations agree either if they match exactly, or, for instances containing at least one verb, noun, adverb or adjective, if the mismatches only concern components of other parts of speech (e.g., a partial match will be counted for *dans l' ensemble* ‘in the set’ \Rightarrow ‘overall’, whether or not both annotators have included *dans* and/or *l'*).

As can be seen, the agreement for NEs is good, and much higher than for non-verbal MWEs. It is only slightly better in partial match than in exact match, which proves that the disagreement concerns more the MWE status than their span. Given the care taken in designing the guidelines, the obtained agreement is somewhat disappointing.⁴² We found out though that the global agreement score masks differences among the various subcorpora within the Sequoia corpus. Namely, the agreement scores are roughly equivalent for the non-medical subcorpora, but much lower for the medical subcorpus, both

⁴² Nonetheless, MWE annotation quality is rarely evaluated. For instance, in the French Treebank, the quality of MWE annotation was not measured. For the PolyCorp corpus, agreement was computed on the categorization of MWEs only (Tutin and Esperança-Rodier 2019). For English, Schneider *et al.* (2014) report a 65% agreement on a 200-sentences sample, but this is not fully comparable, because their metric is different, their scope considers multiword NEs as MWEs (for which the task is easier), as well as “weak” MWEs (roughly, collocations).

for non-verbal MWEs and for named entities.⁴³ These figures reveal that the task is particularly hard for corpora from a technical domain such as medicine. This is probably due to the fact that establishing the MWE-hood of technical terms requires a domain expertise which the annotators are missing and which calls for external knowledge sources. During adjudication, we clarified the use of the LEX criterion for technical terms, and the coherence checking tool subsequently helped to ensure the coherence of annotations across sentences.

9 INTERACTION WITH SYNTACTIC ANNOTATION

Recall that we annotated MWEs on top of an existing treebank, in which grammatical MWEs were already tagged (using the French Treebank corpus guidelines (Abeillé and Clément 1999–2015)). The Sequoia dependency trees contain both syntactic arcs and arcs dedicated to MWE encoding: a grammatical MWE is represented using a flat structure, in which all but the first component of the MWE are attached to the first component using a specific dependency label.⁴⁴

We performed the MWE and NE annotation using new guidelines, and independently of the pre-existing MWE annotation. After completing our annotations, we modified the dependency trees in order to obtain a coherent interaction between the MWE status and syntactic representation:⁴⁵ the set of annotated MWEs changed and we used a binary distinction between syntactically irregular MWEs and syntactically regular MWEs. For the former, we keep the flat representation, while for the latter, we use a regular syntactic structure.

⁴³The lower agreement for the medical subcorpus cannot be explained by a higher frequency of annotations: Section 10 shows that there is roughly one annotation (NE or MWE) every 10 tokens in the whole corpus, (Table 3), but one every 14.5 tokens for the medical subcorpus (Table 5).

⁴⁴In the original Sequoia annotation, MWEs were merged tokens. Subsequent versions used the flat representation proposed in SPMRL 2013 (Seddah *et al.* 2013) and equivalent to the Universal Dependencies fixed label.

⁴⁵This is done in agreement with the authors of the Sequoia treebank, and is integrated in Sequoia releases, from 9.0 version onwards.

Distinguishing between syntactically regular
and irregular MWEs

9.1

While irregularity of an MWE may show at various linguistic levels (morphological, syntactic, lexical, ...), most MWEs are syntactically regular but irregular in other respects. This is often the case for MWEs identified as such due to lexical paradigmatic irregularities (namely passing the [LEX] test). For example, in *appel d'offres* 'call of offers' ⇒ 'call for tenders', the irregularities are the unexpected change of meaning when substituting *offres* by a synonym, the impossibility to insert modifiers normally allowed for this pattern, and the frozen plural of the noun *offres*. Yet, the syntactic distribution of the sequence is exactly as expected for a noun modified by a prepositional phrase.

As previously proposed by Candito and Constant (2014) for parsing experiments, we distinguish between *syntactically irregular* and *syntactically regular MWEs*, and for the latter, we disconnect the marking of the MWE status from the morphosyntactic representation. More precisely, we classify an MWE as syntactically regular whenever its external syntactic distribution can be predicted given the sequence of parts of speech of its components. By syntactic we mean that the distribution is tested focusing on grammaticality only, independently of interpretability, and by external distribution, we mean the categories of heads that the MWE can be attached to.⁴⁶ Note that this does not mean that the MWE exhibits full syntactic regularity, in particular the internal modification of the MWE is generally more constrained than for non-MWE sequences.

By definition, a syntactically regular MWE (i) can be represented using a regular internal syntactic structure, otherwise its distribution would not be predictable, and (ii) does not require a part of speech for the whole sequence, since the parts of speech of the components are sufficient to predict the MWE's external distribution.

⁴⁶ Recall (cf. the [IRREG] test on p. 470) that some MWEs exhibit internal regularity but do not have a predictable external distribution, such as *à(-)coup* 'at-shot' ⇒ 'judder', for which the preposition plus noun sequence has the unexpected distribution of a noun. These cases are considered syntactically irregular.

Table 2:
Syntactically regular vs. irregular
annotated MWEs

	Tokens	Types
REGULAR	2,764	1,253
IRREGULAR	687	173
TOTAL	3,451	1,426

All NEs are currently systematically represented using regular syntax. For the verbal MWEs, which were for the most part inherited from the PARSEME project, they all have the external distribution of a verb, verb phrase, or clause (as required by the PARSEME guidelines). For the vast majority of cases, the internal structure is also regular. For instance, all light-verb constructions and inherently reflexive verbs are, by definition, regular.⁴⁷

We can see in Table 2 that, among the non-NE MWEs, approximately one fifth of the occurrences are irregular, but they correspond to approximately 12% of the lexicon of annotated MWEs (169 among 1,423 types, with types defined as ordered sequences of lemmas).

9.2

Part of speech for syntactically irregular MWEs

For a syntactically irregular MWE, by definition, the distribution cannot be regularly determined by the structure of the MWE, so an explicit part of speech is needed to indicate the distribution class of the MWE. We manually assigned the part of speech for irregular MWEs by looking for the POS matching best its distribution.

Special care was taken to distinguish prepositions from adverbs. We tag as prepositions only the MWEs allowing a direct nominal complement (potentially optional). For instance we tag *étant donné* ‘being given’⇒‘given’ as a preposition because it introduces a direct NP (*Étant donnés les résultats,...* ‘Given the results,...’) or a clause. This led us to use the adverb POS for MWEs taking a non direct nominal

⁴⁷The only two borderline cases found are *plaider (non) coupable* ‘plead (non) guilty’ (in which the adjective could be analyzed as a predicative complement, but it is not normally subcategorized for by the verb *plaider* ‘plead’), and *tourner court* ‘turn short’⇒‘come to a sudden end’, in which the use of the adjective is difficult to characterize, although it can be used in the same manner in other contexts such as *Il joue trop court* ‘He plays too short’.

complement, even when the PP complement is mandatory (e.g., *à partir de lundi* ‘at leave of Monday’⇒‘starting from Monday’), although this is not typical for single-word adverbs.

*Automatic modification
of the dependency representations*

9.3

A single annotator classified the annotated MWEs into syntactically regular vs. irregular, first using a classification based on the POS pattern and then manually checking the MWEs for some of the patterns. While some patterns are always regular (e.g., NOUN + ADP + NOUN), others are mixed. For instance *en partie* ‘in part’⇒‘partly’ is regular, but *à travers* ‘at side’⇒‘across’ functions as a preposition, which is not regular for an ADP + NOUN pattern. All MWEs with cranberry words were considered irregular. We then automatically modified the syntactic representation when needed (to turn the dependency representation either into a regular syntactic structure or into a flat representation for irregular MWEs).

The regular vs. irregular distinction cuts across the functional versus lexical MWE distinction. For instance, in (15), *110 mètres haies* ‘110 meters hurdles’ has the distribution of a noun and is irregular (the pattern would rather function as a cardinal + noun combination, blocking the possibility to use another cardinal). On the contrary, *au cours* ‘at-the course’⇒‘during’ has a regular behavior for a preposition plus noun expecting a PP complement (with a required preposition *de* ‘of’). For this latter case, the pre-existing MWE annotation considered *au cours de* as a grammatical MWE tagged as a preposition. We recreated a regular PP dependency structure as shown in Figure 3.⁴⁸

- (15) **Au cours** de sa carrière, elle a remporté deux **110**
 at-the course of her career, she has won two **110**
mètres haies.
 meters hurdles
 ‘During her career, she has won two 110 meters hurdles.’

⁴⁸See the annotation format page: <https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Corpus-format-description>.

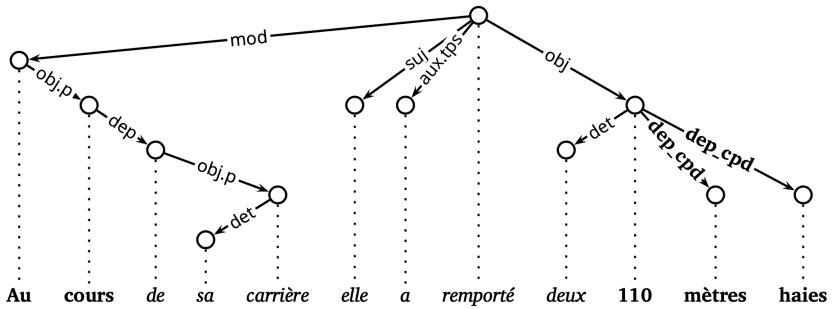


Figure 3: Dependency tree for sentence (15), with one regular MWE (left, bold) and one irregular MWE (right, bold)

Only a few cases show regular internal structure but irregular external distribution (and thus were tagged as irregular). This is the case of *le temps de* ‘the time of’ \Rightarrow ‘by the time’, exemplified in (16).

- (16) **Le temps de** se garer, le magasin était fermé.
 the time of self park, the shop was closed
 ‘By the time we parked, the shop was closed.’

10

CORPUS STATISTICS

The Sequoia corpus comprises 3,099 sentences. The statistics of the final MWE/NE annotation layer, summarized in Table 3,⁴⁹ show that there are 6,579 annotated MWEs/NEs.⁵⁰

Annotations occur at a rate of one MWE/NE every 10.5 tokens. Overall, 11.2% and 7.9% of the tokens belong to MWEs and NEs, respectively, 18.9% belong to any of these two categories, and 0.2% belong to both an MWE and an NE.

⁴⁹ The columns contain: the overall number of annotations (#), the tokens/annotations rate, the discontinuities’ ratio (disc), the average length (len), the average ratio of unseen, seen as variant (var) and identical to seen (ident). The last three values use 10-fold cross-validation, as explained in Section 10.2.

⁵⁰ Additionally, there exist 152 annotations with a primitive NE category (co-existing with another effective NE category for the same tokens, cf. Section 5.1). We disregard them in the following counts.

	#	rate	disc (%)	len	unseen (%)	var (%)	ident (%)
All	6,579	10.5	9.7	2.10	28.2	7.7	64.1
NEs	3,128	22.0	0.4	1.83	30.7	1.9	67.4
MWEs	3,451	19.9	18.1	2.34	26.0	12.8	61.2
↳REG	2,764	24.9	22.3	2.42	29.2	14.1	56.7
↳Verbal	981	70.1	50.6	2.29	37.9	29.3	32.8
↳Others	1,783	38.6	6.7	2.49	24.2	7.1	68.7
↳IRREG	687	100.1	1.0	2.02	13.5	6.1	80.3

Table 3:
Corpus statistics

About half (47.5%) of the annotated instances are NEs, which occur every 22.0 tokens on average, 99.6% of them are continuous, 56.4% are of length one (1845) and they have an overall average length of 1.83 tokens. About 8% of NEs are nested and only 6 of them (0.2%) are overlapping (Section 5.2), as in *Jeanine*_{PERS₁} and *Willy*_{PERS₂} *Schaer*_{PERS₁,PERS₂}.

MWEs account for 52.5% of the annotated entities, and are mostly syntactically regular. About one third of them are VMWEs (inherited from the PARSEME corpus). A VMWE occurs once every 70.1 tokens, with an average length of 2.29 tokens. VMWEs are much more often discontinuous than other categories (50.6% of the time), with an average gap of 0.9 tokens (65% of the discontinuities have a 1-token gap, 20% have a 2-token gap, and so on, up to one MWE containing a 20-token gap). Only 4 and 39 VMWEs (0.4% and 4%) are overlapping and nested, respectively. Examples of the latter include light-verb constructions in which the verb is itself a VMWE, as in *faire*_{1,2} *l*_{1,2} *'*_{1,2} *objet*_{1,2} *d*₂ *'une enquête*₂ ‘make the object of an investigation’ ‘come under investigation’.

Non-verbal MWEs correspond to 37.5% of all annotations, and occur at a rate of 0.8 per sentence (and one non-verbal MWE every 27.8 tokens). They have an average length of 2.36 tokens but, differently from VMWEs, they are mostly continuous (94.9% of the time). Nesting is rare (1.7%), while overlapping is a bit more frequent (4.7% of the non-verbal MWEs share one component with another one). Most non-verbal MWEs are syntactically regular (72.2%). They occur once every 38.6 tokens, have the largest average length (2.49), and 6.7% of them only are discontinuous.

Only 687 MWEs (all non-verbal) are tagged as syntactically irregular. These include all MWEs with a cranberry word. They are almost always continuous (99%) and most of them behave as an adverb (30%, e.g., *aujourd'hui* ‘today’, *peut-être* ‘maybe’ and *bien sûr* ‘of course’) or preposition (27%, e.g., *en tant que* ‘as’ or *suite à* ‘after’). The partitive determiner *du* (contraction of *de le* ‘of the’) accounts for 5% of all irregular MWEs.

10.1 *Frequency of use of the MWE sufficient criteria*

Recall from Section 8.2 that, for non-verbal MWEs, the guidelines provide a list of sufficient criteria, among which annotators had to provide only one, although several criteria may apply. During adjudication, one of the two provided criteria was randomly chosen. This makes it possible to compute statistics of how often each criterion was used. The LEX criterion, which targets the limited paradigmatic variability, is by far the most frequent (1544 times), followed by IRREG (361), DET (210), CRAN (155), INSERT (74), ZERO (46), SYNT (33), and MORPHO (32). The two semantic criteria PRED and ID were used only 8 and 4 times. Note that the LEX criterion is not very formal, in the sense that the annotator is asked to evaluate the unexpectedness of a meaning shift. This might provide an explanation for the medium level of inter-annotator agreement for MWEs.

10.2 *Variability*

To estimate the variability of the annotated MWEs/NEs, we used a method inspired by 10-fold validation.⁵¹ In each turn, we defined as seen those MWEs/NEs which were annotated in the 9 training folds. By an identical annotation (*ident*) we mean an MWE/NE which had the same sequence of word forms, with the same gap lengths, as a seen MWE/NE. By a variant annotation (*var*) we understand a non-identical annotation sharing its multiset of lemmas with a seen MWE/NE. Finally, an annotation is defined as unseen if it shares its multiset of lemmas with no seen MWE/NEs.

⁵¹ One fold was made of one sentence every ten sentences, hence the folds covered the four subcorpora with the same proportions as the whole corpus.

	# annotations	# distinct annotations	Ratio
NEs	3,128	1,401	2.23
MWEs	3,451	1426	2.42
↳IRREG	687	173	3.97
↳Verbal REG	981	524	1.87
↳Non-Verbal REG	1,783	729	2.44

Table 4:
Annotation
variability
statistics

These ratios are shown in the last three columns of Table 3. VMWEs exhibit the highest variability, having both the highest unseen ratio (37.9) and variant ratio (29.3), and thus the lowest ratio of identical occurrences. All the other kinds of annotations (NEs, non-verbal regular MWEs and non-verbal irregular MWEs) have a much lower variant ratio (1.9%, 7.1% and 6.1% respectively). NEs also have a high unseen ratio (30.7%), but since they exhibit very low morphosyntactic variability (1.9%), they also have a high ratio of identical occurrences (67.4%). Irregular MWEs have the lowest variability: on average, only 13.5% of them were not seen and, when seen, they were generally identical (80.3% of the time).

The token/type ratio, that is, the average number of annotations per multiset of lemmas, is another variability indicator in the annotated MWEs/NEs. The lower the ratio, shown in the last column of Table 4,⁵² the less frequently entities re-occur and thus the higher their variety. Surprisingly, the less varied category are the irregular MWEs, with an average number of 3.97 tokens per type, while the verbal MWEs are the more varied (1.87 tokens per type).

Breakdown by subcorpus

10.3

The statistics in the four subcorpora are shown in Table 5. The overall density of annotations (inverse of the rate column) is comparable for Europarl and regional news, a little higher for the Wikipedia narrative texts, and – interestingly – lower for medical reports. Divergences occur across categories: NEs are frequent in Wikipedia (one every 12.7 tokens), and rather rare in medical reports (one every 49.1 tokens, mainly corresponding to drug names). VMWEs are almost twice more

⁵²Two annotations are *distinct* if their multisets of lemmas differ.

Table 5:
Statistics
broken down
by subcorpus;
the column
headers
are defined
as in Table 3

	#	rate	disc (%)	len	unseen (%)	var (%)	ident (%)
All (MWEs and NEs)							
Regional news	1,144	10.0	10.3	2.0	57.6	6.1	36.3
Europarl	1,361	11.3	13.2	2.1	33.5	8.9	57.5
French wiki	2,724	8.3	5.0	2.2	33.5	6.6	59.9
Medical reports	1,350	14.5	15.0	2.0	14.7	7.1	78.2
NEs							
Regional news	519	21.9	0.6	1.8	57.8	0.8	41.5
Europarl	437	35.1	0.2	1.5	31.3	0.2	68.5
French wiki	1,773	12.7	0.6	2.1	30.5	2.9	66.7
Medical reports	399	49.1	0.0	1.1	6.0	0.0	94.0
Verbal Regular MWEs							
Regional news	204	55.8	44.6	2.3	73.0	20.1	6.9
Europarl	295	52.0	45.4	2.3	47.5	26.1	26.4
French wiki	221	101.6	38.9	2.4	53.4	26.2	20.4
Medical reports	261	75.1	70.9	2.2	34.9	19.9	45.2
Non-Verbal Regular MWEs							
Regional news	268	42.5	7.8	2.5	55.0	6.5	38.5
Europarl	388	39.5	10.6	2.5	33.2	5.2	61.6
French wiki	590	38.0	6.8	2.6	33.9	9.5	56.6
Medical reports	537	36.5	3.4	2.4	14.3	6.0	79.7
Irregular MWEs							
Regional news	153	74.4	2.0	1.8	33.3	13.7	52.9
Europarl	241	63.6	1.2	2.2	16.6	8.7	74.7
French wiki	140	160.4	0.7	1.9	33.6	12.1	54.3
Medical reports	153	128.0	0.0	2.0	14.4	6.5	79.1

frequent in regional news and Europarl than in Wikipedia narratives. The frequency of non-verbal regular MWEs does not vary much across subcorpora, although they are slightly more frequent in the medical subcorpus.

The variability of annotations is the more spread-out property across subcorpora. For all the categories of annotations, we observe the highest unseen ratio in the regional news subcorpus, an interme-

diate ratio for the Europarl and Wikipedia subcorpora, and the lowest ratio for medical reports. This can be explained more by the number of documents contained in each subcorpus than by genre differences across subcorpora: the medical subcorpus consists of two reports concerning marketing authorization for two specific drugs, with a very focused topic. Conversely, the regional news concern very varied topics (which may explain the high unseen ratio of NEs: 57.8%), and the Wikipedia corpus contains about 20 Wikipedia narrative pages.

The proportion of discontinuous MWEs or NEs is stable across subcorpora, except a high ratio of discontinuous verbal MWEs in the medical reports (70.9%), due to a higher proportion of light-verb constructions, which tend to be discontinuous.

We provide the most frequent MWEs/NEs in Table 6, for each subcorpus. For each subcorpus, its most frequent NEs are specific to its topics (for instance specific drugs for the medical reports, or European institutions for Europarl). Verbal MWEs also reveal the domain of some of the subcorpora, in particular we observe legal vocabulary for the French Wiki subcorpus, which relates famous contemporary politico-financial affairs. For irregular MWEs, the only feasible observation is that, in Europarl, these are more argumentative or formal.

Comparison to other corpora

10.4

The closest feasible comparison that we can draw is that between our annotations on the Sequoia treebank, and the MWE annotation of the French Treebank (FTB), which include multitoken named entities. Note that the two corpora have quite different sizes (about 650k tokens for FTB, and about 70k tokens for Sequoia), and genres partially match (FTB is mono-genre with sentences from *Le Monde*, versus four genres for the Sequoia). The MWE annotation process is also different: FTB being larger, MWEs were automatically pre-identified and then manually annotated in a mono-annotator setting. For Sequoia, we used no pre-annotation tool (to avoid bias) and performed double-annotation, making it possible to compute inter-annotator agreement.

Nevertheless, the density of the annotated MWEs and NEs turns out to be similar, provided some cases annotated in Sequoia only are ignored. More precisely, when setting aside our annotated single-token named entities, we have 4,830 MWEs/NEs, occurring at a rate

Table 6:
Most frequent
MWEs/NEs
for each
subcorpus
(for better
readability,
for some
of the examples
we provide
the most
frequent
inflected form)

Most frequent cases	
NEs	
Reg. news	France, Belfort, Montbéliard
Europarl	Commission, Parlement ('Parliament'), Union Européenne 'European Union')
French wiki	Paris, RPR, Taïwan
Med. reports	Aclasta, Angiox, Paget
Verbal Regular MWEs	
Reg. news	<i>il faut</i> 'it-EXPL must'⇒'it is necessary/mandatory to', <i>se dérouler</i> 'REFL unfold'⇒'to happen', <i>il s'agit</i> 'it REFL acts'⇒'it is / it is about'
Europarl	<i>il faut</i> 'it-EXPL must'⇒'it is necessary/mandatory to', <i>il s'agit</i> 'it REFL acts'⇒'it is / it is about', <i>il y a</i> 'it there have'⇒'there is'
French wiki	<i>mettre en examen</i> 'place under formal investigation', <i>il y a</i> 'it there have'⇒'there is', <i>il s'agit</i> 'it REFL acts'⇒'it is / it is about', <i>avoir lieu</i> 'have place'⇒'to take place', <i>mettre en cause</i> 'put into cause'⇒'implicate'
Med. reports	<i>se produire</i> 'REFL produce'⇒'to happen', <i>atteint d'insuffisance</i> 'affected by insufficiency', <i>avoir fracture</i> 'have fracture'
Non-Verbal Regular MWEs	
Reg. news	<i>à l'occasion</i> 'at the occasion', <i>jeune fille</i> 'young girl', <i>dans un Xième temps</i> 'in a x-th time'⇒'over a x-th phase'
Europarl	<i>Etat membre</i> 'member state', <i>droits de l'homme</i> 'rights of the man'⇒'human rights', <i>dans le cadre de</i> 'in the frame of'⇒'as part of'
French wiki	<i>marché public</i> 'marker public'⇒'public contract', <i>à l'époque</i> 'at the time', <i>abus de biens sociaux</i> 'misuse of corporate assets'
Med. reports	<i>acide zolédronique</i> 'zoledronic acid', <i>maladie de Paget</i> 'Paget's disease', <i>en cas de</i> 'in case of'
Irregular MWEs (other than cranberry or typographic characters)	
Reg. news	<i>grâce à</i> 'grace to'⇒'thanks to', <i>du</i> (= <i>de</i> + <i>le</i>) 'of the', <i>il y a</i> 'it-EXPL there is'⇒'ago'
Europarl	<i>en tant que</i> 'in so-much that'⇒'as', <i>c'est pourquoi</i> 'this is why', <i>en ce qui concerne</i> 'in what that concerns'⇒'concerning'
French wiki	<i>ainsi que</i> 'so that'⇒'as well as', <i>grâce à</i> 'grace to'⇒'thanks to', <i>à partir de</i> 'to leave from'⇒'starting from'
Med. reports	<i>à moins</i> 'to less'⇒'unless', <i>du</i> (= <i>de</i> + <i>le</i>) 'of the', <i>y compris</i> 'there included'⇒'including'

of one MWE/NE every 14.2 tokens on average. When computing this rate in FTB (using its dependency 1.0 version), we obtain a slightly lower density: the rate is one MWE/NE every 18.7 tokens, and one every 21.0 tokens when ignoring the numerical MWEs in FTB.

To compare the density for MWEs other than named entities, we can set aside the MWEs annotated in FTB that are tagged as proper nouns. We obtain 25,656 MWE annotations in FTB, both non numerical and not tagged as proper nouns, occurring at a rate of one every 25.1 tokens, compared to 19.9 for the (non-NE) MWEs in Sequoia. The lower MWE density in FTB is explainable by the scarcity of discontinuous MWEs, which have limited the annotation of verbal MWEs. When ignoring the discontinuous MWEs in Sequoia, we obtain a rate of one continuous MWE every 24.3 tokens, hence quite similar to that of FTB.

When comparing the FrWiki part of Sequoia to the English Wiki50 corpus (Vincze *et al.* 2011), we find the same rate for NEs (one NE every 12.8 tokens), but a slightly lower density for MWEs (one MWE every 29.7 tokens in Wiki50, versus one MWE every 23.6 tokens for the FrWiki part of Sequoia). This may be explained by the absence of functional MWEs in Wiki50. Another important resource for English, the Streusle corpus, contains 3,013 “strong MWEs” (including some NEs) and 705 “weak MWEs” (collocations, disregarded in Sequoia), yielding a density of one strong MWE/NE every 18.4 tokens. This is slightly lower than the one in Sequoia (one MWE/NE every 14.2 tokens), and similar to FTB (one MWE every 18.7 tokens). These figures remain hard to compare, though, given the corpora’s different annotation scopes.

FINDINGS

11

Let us mention several lessons learned from this endeavor. Firstly, we initially intended not to differentiate between NEs and MWEs, but to include the annotation of multiword NEs as nominal MWEs. Such an approach might, in particular, solve the problem of heterogeneous typologies (cf. Section 3.3). It would also make the annotation flowcharts simpler because some tests could be shared, notably between terminological MWEs (whose annotation often requires expert

knowledge) and NEs having a descriptive basis (i.e. not pure proper nouns). However, such an integration proved hard to achieve. As an alternative, we proposed an NE-specific decision tree, holding for all types of NEs, and capitalizing on the specificity of the naming convention existing for NEs. We leave it as future work to test a unified modeling principle.

Another heterogeneity issue stems from the fact that verbal MWE annotation follows a detailed flowchart with about 40 (mostly subcategory-dependent) tests, while non-verbal MWEs are all contained in one category and covered by 10 generic tests, each of which is considered individually sufficient. For the non-verbal parts-of-speech, we had the objective to propose a simple and generic list of sufficient criteria. In the end, we proved that 10 generic criteria are indeed sufficient to cover all non-verbal MWEs, and achieve substantial inter-annotator agreement. Importantly, we hypothesize that these criteria are more portable to other languages.

Another finding has been the relative hardness of capturing functional MWEs. Many previous efforts towards modeling and annotating MWEs started with multiword prepositions, conjunctions, pronouns, and other functional bundles, considered easier to capture due to their contiguity and morphosyntactic inflexibility. Conversely, we found that when functional MWEs lack an open-class component (e.g. in *d'entre*, *d'après*), the lexical substitution (LEX), identity (ID), predication (PRED), determiner fixedness (DET) or inexistence (ZERO) and morphosyntactic fixedness (MORPHO, SENT, INSERT) criteria can hardly be used. When a functional MWE does contain open-class components, such components have a very general meaning, or lack possible substitutes needed to test the LEX criteria (*dans le cadre de* 'in the frame of' ⇒ 'as part of'). The criteria for testing fixedness are used instead in this case (MORPHO, INSERT). Additionally, closed-class parts-of-speech often mask fine-grained distribution distinctions (for instance prepositions allowing a determiner-less NP or not, tested by the ZERO criterion, as in *en parallèle* 'in parallel' ⇒ 'simultaneously').

CONCLUSIONS AND FUTURE WORK

12

We presented the annotation of named entities and multiword expressions in Sequoia (Candito and Seddah 2012), a French treebank covering various written genres (news, parliamentary debates, wikipedia narratives, and medical reports). The corpus comprises 3,099 sentences, in which we annotated 3,112 NEs and 2,459 non-verbal MWEs. These complement the 981 verbal MWEs previously annotated on the same data within the COST PARSEME project. Although rather modest in size, the resulting corpus is the only open-source treebank for French annotated with MWEs and NEs.

A contribution of this work is that our MWE/NE typology is endorsed by extensive annotation guidelines based on decision flowcharts over linguistic tests, which are meant to guide the annotator – in a relatively deterministic and reproducible way – to both identify and categorize candidates into one of the proposed categories. In particular, we largely cover the challenge of distinguishing NEs and MWEs, in terms of operational definitions and in the presence of intimate interactions between these phenomena. To the best of our knowledge, this constitutes an unprecedented outcome.

Moreover, a fundamental trait of our approach is to model the MWE status separately from the syntactic annotation: depending on its distribution and internal pattern, a given MWE can be considered regular from the syntactic point of view, and hence receive a regular internal structure. Another originality stands in our choice to use sufficient criteria for the MWE status. Namely, various combinations of idiomaticity criteria may or may not apply to various MWEs, which results in a high variety of idiomaticity profiles. It would be very challenging to quantify this variability, and especially to establish an objective threshold above which a candidate proves idiomatic enough to be considered an MWE. We avoid this difficulty by considering that fulfilling any of the (sufficient) criteria is enough for a candidate to be marked as an MWE.

The resulting resource thus comprises annotated MWEs with varying degree of idiosyncrasy. One possible future extension concerns characterizing the degree of compositionality of the annotated MWEs, for instance, by estimating the semantic contribution of each compo-

ment to the whole MWE. Another interesting research question would be to what extent our annotation guidelines, covering NEs and all categories of MWEs, could scale up to many languages, just as the multilingual PARSEME guidelines for verbal MWEs do. We hope that this resource will enable research both in linguistic modeling and automatic identification methods which can jointly deal with NEs, verbal MWEs, and non-verbal MWEs.

APPENDIX: IDENTIFICATION CRITERIA FOR NON-VERBAL MWES

Semantic identity [ID]: Semantic criteria are tricky because they rely on less formalized notions than lexical and syntactic criteria. Therefore, we restrict their application to nominal expressions, for which two simple tests help to signal that one of the content words has an unusual meaning. Following Gross (1988), the semantic criterion ID checks whether *c* is a hyponym of its syntactic head *h*. If this is not the case, the test confirms that, in the context of *c*, the head *h* does not have one of its usual senses. In practice, we systematically test whether “*a c is a h*” is semantically acceptable. If not, *c* is annotated as MWE. The test passes, for instance, for *cordons bleus* ‘excellent cook’, which is not a *cordons* ‘cord’.

Predicative relation [PRED]: In the case of noun-adjective candidates, a second semantic test concerns the predicative relation between the adjective *a* and the noun *n*. If the adjective⁵³ cannot be used in a predicative construction with the noun *n*, then the candidate is a MWE, as illustrated in (17).

- (17) #L’ arme blanche est blanche.
the weapon white is white
‘The cold weapon is white.’

⁵³The test only applies for adjectives that can be used in predicative mode.

Cranberry word [CRAN]: A component of *c* does not function as an isolated word, and can only be used in a very restricted number of combinations, usually one or two. For instance, the words *catimini* and *tandis* in the expressions *en catimini* ‘on the quiet’ and *tandis que* ‘whereas’ are used in these expressions only. The word *afin* cannot be used but in the complex preposition *afin de* ‘in order to’ or in the complementizer *afin que* ‘so that’.

Limited lexical substitution [LEX]: A standard criterion to capture semantic idiomaticity is to test the impossibility of substituting content words (i.e., nouns, verbs, adjectives and adverbs) in *c* by semantic neighbors, namely synonyms, antonyms, or hypernyms. More precisely, applying such a substitution would produce either a forbidden combination or a combination whose meaning shift goes beyond the expected initial substitution. For instance, going from *eau sucrée* ‘water sweet’ ⇒ ‘sweet water’ to *boisson sucrée* ‘drink sweet’ ⇒ ‘sweet drink’, the meaning shift between *eau* ‘water’ and *boisson* ‘drink’ is encompassed within the meaning shift between *eau sucrée* and *boisson sucrée*. However, when transforming *eau de vie* ‘water of life’ ⇒ ‘brandy’ into *boisson de vie* ‘drink of life’, the meaning shift is greater than the one between *eau* and *boisson*. Example (18) shows another case of unexpected meaning shift and unacceptable modification for a candidate containing a single content word:

- (18) à la (suite | #succession | *continuité) de
to the (following | #succession | *continuity) of
‘following’

This criterion also applies for technical or institutional multiword terms, if the domain specificity is lost when substituting one component. For instance, when moving from *juge d’instruction* ‘judge of investigation’ ⇒ ‘examining magistrate’ to *juge d’investigation* ‘judge of investigation’ we retain the general meaning, but lose the precise meaning of a specific profession in the French judiciary system. We thus annotate as MWE all candidates referring to institutional professions. We also use this criterion for technical terms, for which we know⁵⁴ that they name a precise technical concept whose formula-

⁵⁴ Or we can check using external specialized lexical resources.

tion is frozen and comprises a surplus of meaning with respect to the composition of its parts. For instance, in *traduction automatique* ‘translation automatic’ \Rightarrow ‘machine translation’, switching to *traduction automatisée* ‘translation automatised’ is understandable, but does not refer to the technical domain of machine translation anymore.

As shown in the corpus statistics (Section 10.1), the LEX criterion is by far the most frequently used. A posteriori, it would have been more informative to split it according to the kind of unexpected meaning shift obtained when substituting one component.

We also use this criterion for multiword names of artefact models or brands, when they are used to refer to instances of such a model or brand. For instance in (19), *Rolls Royce* refers to one specific organization and is tagged as a NE, whereas in example (20), it refers to a specific car. The naming convention here applies for any car of the Rolls Royce brand, hence it is not a NE (the outcome of the CONCEPT_NAMING_CONV test in the top decision flowchart of Section 4.2 is YES, redirecting to the non-verbal MWE guide).

- (19) [Rolls Royce]_{ORG} a annoncé son bénéfice 2018.
Rolls Royce has announced its profit 2018
‘Rolls Royce has announced its 2018 profit.’
- (20) J’ ai acheté une (Peugeot 308 | Rolls Royce).
I have bought a (Peugeot 308 | Rolls Royce)

Fixed determiner [DET]: If the determiner of a noun appearing in *c* is totally frozen, except for number or gender variation, it suffices to identify the candidate as a MWE. Note that we include as a special case of fixed determiner the case of a fixed “zero” determiner, that is, when a determiner is impossible whereas there should normally be a determiner according to general grammar. However, there are several productive contexts in which a noun can occur without a determiner, so the guidelines list cases (not detailed here) for which the absence of determiner should not be considered as a sufficient criterion for MWE identification. Also note that we distinguish between a fixed zero determiner (which we include in this criterion as a special case of a fixed determiner), and the unexpected possibility to have a zero determiner (criterion ZERO).

When the determiner is fixed under certain conditions only, we do not consider the test passed. In particular, the determiner can be frozen when the noun has no modifier (as *après-midi* ‘afternoon’ in (21)), but more variable otherwise (as in (22)).

- (21) à cinq heures de l’ après-midi
 at five hours of the afternoon
 ‘at five p.m.’
- (22) à cinq heures d’ une après-midi (*∅ | de juillet)
 at five hours of an afternoon (*∅ | of July)
 ‘at five o’clock on a July afternoon’

Moreover, we apply specific tests for candidates that include a noun phrase (NP) introduced by the preposition *de* ‘of’, that is, following the pattern ADP + [DET]⁵⁵ + NOUN + *de* + NP such as *à l’origine du problème* ‘at the’origin of the problem’. We consider that the determiner is *not* fixed when the *de* + NP sequence can be replaced by the interrogative determiner *quel* ‘what’,⁵⁶ as in examples (23) and (24) (Danlos 1980).⁵⁷

- (23) en l’ honneur de la République
 in the honor of the Republic
- (24) En quel honneur est donné ce banquet?
 in what honor is given this banquet
 ‘In what honor this banquet is given?’

Conversely, the test is passed if the determiner, otherwise fixed, alternates with a possessive determiner whose antecedent is the (un-expressed) *de* + NP, as in example (25). Note that in such cases, we consider that the DET criterion is sufficient to tag the sequence as a MWE, but the determiner is not included in it, to homogenize annotation for the two instances of the same MWE in (25).

- (25) à la recherche du Graal / à sa recherche
 at the search of-the Graal / at its search
 ‘in search of the Graal / in search of it’

⁵⁵The determiner is optional.

⁵⁶We thank Laurence Danlos who suggested this test to us.

⁵⁷The applicability of the test has some restrictions, e.g., it does not apply if the NP is animated, because *quel* ‘what’ never refers to animated entities.

Possible absence of determiner [ZERO]: This criterion is satisfied whenever the determiner can be both present and absent, in a pattern that normally requires a determiner, as in (26). As for the previous criteria, we ignore the regular cases of zero determiner. For instance, certain prepositions such as *avec* ‘with’, *pour* ‘for’, and *sans* ‘without’ can introduce NPs without determiners.

- (26) à (∅ | son) domicile
 at (∅ | his-or-her) home

Limited morphological variation [MORPHO]: A MWE can be identified whenever a given regular morphosyntactic rule fails for *c*, according to general grammar. This comprises morphological features (e.g., tense, number, gender) and analytic verbal tenses and moods. Either a given form is impossible, as in (27), or agreement is breached (e.g., *un peau rouge* ‘a.MASC skin.FEM red’⇒‘a redskin’).

- (27) un (garde du corps | #garde des corps)
 a (guard of.the.SG body | #guard of.the.PL bodies)
 ‘a bodyguard’

Irregular morphosyntactic structure [IRREG]: If *c* shows an irregular morphosyntactic structure, its global meaning cannot be derived using compositional operations, and we tag it as a MWE. The irregularity can stem from the internal structure or the external distribution.

For the internal (ir)regularity, the test evaluates whether the combination of components of such parts of speech is regular, independently of semantics. For instance *à peu près* ‘at little close’⇒‘approximately’ combines a preposition introducing an adverb, which is not regular. For closed grammatical categories, the test sometimes considers the components and not just their category. For instance the sequence *en outre* ‘in besides’⇒‘in addition’ is the juxtaposition of two prepositions, which is not regular for the preposition *en*.

The test also passes when the internal structure is regular, but it does not have the expected *external* distribution. For instance, the sequence *longue portée* ‘long range’⇒‘long-range’ is regularly composed of an adjective modifying a noun, but has the distribution of a postnominal adjective, unexpected in French for such a combination.

- (28) Le suspect est armé d' un fusil longue portée.
the suspect is armed of a rifle long range
'The suspect is armed with a long-range rifle.'

This test also passes for certain adverb + *que* sequences (Section 12).

Limited syntactic variation [SYNT]: We annotate a candidate *c* as a MWE whenever morphosyntactic variations that should apply, given the candidate's morphosyntactic pattern, are not possible for *c*.

This criterion covers three specific nominal patterns. The first pattern is NOUN₁ + ADJ, which usually accepts the variation NOUN₁ *de* 'of' [DET] NOUN₂, with NOUN₂ morphologically related to the ADJ (e.g., a denominal adjective). For instance, *produit régional* 'regional product' is synonym to *produit de la région* 'product of the region'. This alternation is not possible, however, for *conseil régional* 'regional council' vs. *#conseil de la région* 'council of the region', which designates the legislature of a French region (political division). Thus, *conseil régional* is a MWE according to this criterion.

The second pattern is NOUN₁-NOUN₂ (two nouns linked by a hyphen). Regularly, the order of the nouns is arbitrary (e.g., *plombier-serrurier* 'plumber-locksmith' is equivalent to *serrurier-plombier* 'locksmith-plumber'). When this is not possible, the criterion indicates a MWE (e.g., *sapeur-pompier* 'sapper-firefighter' ⇒ 'firefighter' but not **pompier-sapeur*). Nonetheless, the criterion cannot be used when the meaning change is productive and predictable, such as in *le trajet Paris-Strasbourg* 'the Paris-Strasbourg route'.

The third pattern concerns the shift from prenominal to postnominal position for adjectives which can be regularly postposed. It is almost exclusively applied to *jeune (homme | femme)* 'young (man | woman)'. Postposition induces a slight meaning shift, with more focus on the age of the person.

Limited insertion [INSERT]: This criterion tests for the insertion of material that is, in theory, syntactically compatible and semantically plausible for one of the candidate components.⁵⁸ This regular insertion is not possible for MWEs, as shown in examples (29)–(30):

⁵⁸ For this test we exclude the use of modifiers such as *dit* 'said' or *soi-disant* 'self-saying' ⇒ 'supposed', which have a metalinguistic meaning.

- (29) Le processus est **en cours** (*normal).
the process is in course (*normal)
'The process is ongoing.'
- (30) À l'**issue** (*inattendue) du discours, il est parti.
at the 'exit' (*unexpected) of the speech, he is left
'He left after the speech.'

Particular cases: sequences of the form adverb + que

We found it difficult to decide the MWE status for certain sequences of the form ADV + *que* 'that'. Although usually included in MWE lexicons (Ramisch *et al.* 2016), the number of applicable tests for these is rather reduced. There is a general intuition that the meaning of the adverb is often not present in the ADV + *que*, but this is sometimes difficult to capture given the above tests. For instance, in (31), *alors que* 'then that' ⇒ 'although' has a clear contrastive meaning, which is not present in the meaning of the adverb *alors*. This non compositionality is difficult to capture with the above tests.⁵⁹

- (31) Il a dit rouge **alors que** c'est bleu.
he has said red then that it is blue
'He said red although it is blue.'

We used the IRREG criterion for the ADV + *que* sequences which may function as clause modifiers, namely in "MatrixClause + ADV + *que* + Clause2" contexts. We considered this trait as irregular (IRREG criterion satisfied), given that for almost all adverbs, removing the *que* + Clause2 either leads to unacceptability or modifies the meaning of the adverb (the only exception being *alors* 'then' in its temporal meaning). Note that other adverbs may introduce a *que* + Clause and function as sentence heads, not as clause modifiers. This case is not considered irregular.

⁵⁹ Moreover, several French conjunctions historically formed by an adverb + *que* are now written without separator (e.g., *lorsque* 'when', *puisque* 'since').

REFERENCES

- Anne ABEILLÉ and Lionel CLÉMENT (1999–2015), Corpus le Monde, annotation morpho-syntaxique : Les mots simples – les mots composés, <http://ftb.linguist.univ-paris-diderot.fr/fichiers/public/guide-morphosynt.pdf>.
- Anne ABEILLÉ, Lionel CLÉMENT, and Loïc LIÉGEOIS (2019), Un corpus arboré pour le français : le French Treebank, *Traitement Automatique des Langues*, 60(2):19–43.
- Anne ABEILLÉ, Lionel CLÉMENT, and François TOUSSENEL (2003), Building a treebank for French, in Anne ABEILLÉ, editor, *Treebanks: Building and using parsed corpora*, pp. 165–187, KluwerAcademic Publishers, Dordrecht, The Netherlands.
- Timothy BALDWIN and Su Nam KIM (2010), Multiword expressions, in Nitin INDURKHYA and Fred J. DAMERAU, editors, *Handbook of natural language processing, second edition*, pp. 267–292, CRC Press, Boca Raton.
- Eduard BEJČEK and Pavel STRAŇÁK (2010), Annotation of multiword expressions in the Prague Dependency Treebank, *Language Resources and Evaluation*, 44(1–2):7–21.
- Eduard BEJČEK, Pavel STRAŇÁK, and Daniel ZEMAN (2011), Influence of treebank design on representation of multiword expressions, in Alexander F. GELBUKH, editor, *Proceedings of CICLing 2011 (volume 1)*, pp. 1–14, Tokyo, Japan.
- Conor CAFFERKEY, Deirdre HOGAN, and Josef VAN GENABITH (2007), Multi-word units in treebank-based probabilistic parsing and generation, in *Proceedings of RANLP 2007*, pp. 98–103, Borovets, Bulgaria.
- Nicoletta CALZOLARI, Charles J. FILLMORE, Ralph GRISHMAN, Nancy IDE, Alessandro LENCI, Catherine MACLEOD, and Antonio ZAMPOLLI (2002), Towards best practice for multiword expressions in computational lexicons, in *Proceedings of LREC 2002*, pp. 1934–1940, Las Palmas, Spain.
- Marie CANDITO, Mathieu CONSTANT, Carlos RAMISCH, Agata SAVARY, Yannick PARMENTIER, Caroline PASQUER, and Jean-Yves ANTOINE (2017), Annotation d’expressions polylexicales verbales en français, in *Proceedings of TALN 2017*, pp. 1–9, Orléans, France.
- Marie CANDITO and Matthieu CONSTANT (2014), Strategies for contiguous multiword expression analysis and dependency parsing, in *Proceedings of ACL 2014 (volume 1: long papers)*, pp. 743–753, Baltimore, USA.
- Marie CANDITO and Djamé SEDDAH (2012), Le corpus Sequoia : Annotation syntaxique et exploitation pour l’adaptation d’analyseur par pont lexical, in *Proceedings of JEP-TALN-RECITAL 2012*, pp. 321–344, Grenoble, France.

- Dolors CATALÀ and Jorge BAPTISTA (2007), Spanish adverbial frozen expressions, in *Proceedings of MWE 2007*, pp. 33–40, Prague, Czech Republic.
- Nancy A. CHINCHOR (1997), Appendix E: MUC-7 named entity task definition, in *Proceedings of MUC-7*, Fairfax, USA.
- Nancy A. CHINCHOR (1998), Overview of MUC-7, in *Proceedings MUC-7*, Fairfax, USA.
- Jacob COHEN (1960), A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20:37–46.
- Mathieu CONSTANT, Gülşen ERYIĞIT, Johanna MONTI, Lonneke VAN DER PLAS, Carlos RAMISCH, Michael ROSNER, and Amalia TODIRASCU (2017), Multiword expression processing: A survey, *Computational Linguistics*, 43(4):837–892.
- Ann COPESTAKE, Fabre LAMBEAU, Aline VILLAVICENCIO, Francis BOND, Timothy BALDWIN, Ivan A. SAG, and Dan FLICKINGER (2002), Multiword expressions: linguistic precision and reusability, in *Proceedings of LREC 2002*, pp. 1941–1947, Las Palmas, Spain.
- Laurence DANLOS (1980), *Représentations d'informations linguistiques : constructions N être Prép X*, Ph.D. thesis, Université Paris 7, France.
- Maud EHRMANN (2008), *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*, Ph.D. thesis, Université Paris Diderot, France.
- Karèn FORT and Benoît SAGOT (2010), Influence of pre-annotation on POS-tagged corpus development, in *Proceedings of LAW 2010*, pp. 56–63, Uppsala, Sweden.
- Peter FRECKLETON (1985), Sentence idioms in English, *Working Papers in Linguistics*, 11:153–168.
- Guillaume GRAVIER, Gilles ADDA, Niklas PAULSSON, Matthieu CARRÉ, Aude GIRAUDEL, and Olivier GALIBERT (2012), The ETAPE corpus for the evaluation of speech-based TV content processing in the French language, in *Proceedings of LREC 2012*, pp. 114–118, Istanbul, Turkey.
- Gaston GROSS (1988), Degré de figement des noms composés, *Langages*, 90:57–72.
- Maurice GROSS (1986), Lexicon-grammar: the representation of compound words, in *Proceedings of COLING 1986*, pp. 1–6, Bonn, Germany.
- Maurice GROSS (1994), The lexicon-grammar of a language: application to French, in Ashley R. E., editor, *The encyclopedia of language and linguistics*, pp. 2195–2205, Pergamon Press, Oxford, UK.
- Cyril GROUIN, Sophie ROSSET, Pierre ZWEIGENBAUM, Karèn FORT, Olivier GALIBERT, and Ludovic QUINTARD (2011), Proposal for an extension of

traditional named entities: from guidelines to evaluation, an overview, in *Proceedings of LAW 2011*, pp. 92–100, Portland, USA.

Jan HAJIČ, Eva HAJIČOVÁ, Marie MIKULOVÁ, and Jiří MÍROVSKÝ (2017), Prague Dependency Treebank, *Handbook on linguistic annotation*, pp. 555–594, Springer Handbooks, Springer Verlag, ISBN 978-94-024-0879-9.

Georges KLEIBER (1996), Noms propres et noms communs : un problème de dénomination, *Meta*, 41(4):567–589.

Georges KLEIBER (2001), Remarques sur la dénomination, *Cahiers de Praxématique*, 36:21–41.

Georges KLEIBER (2007), Sur le rôle cognitif des noms propres, *Cahiers de Lexicologie*, 91(2):153–167.

Eric LAPORTE, Takuya NAKAMURA, and Stavroula VOYATZI (2008a), A French corpus annotated for multiword nouns, in *Proceedings of MWE 2008*, pp. 27–30, Marrakech, Morocco.

Éric LAPORTE (2018), Choosing features for classifying multiword expressions, in Manfred SAILER and Stella MARKANTONATOU, editors, *Multiword expressions: insights from a multi-lingual perspective*, pp. 143–186, Language Science Press, Berlin, Germany.

Éric LAPORTE, Takuya NAKAMURA, and Stavroula VOYATZI (2008b), A French corpus annotated for Multiword Expressions with adverbial function, in *Proceedings of LAW 2008*, pp. 48–51, Marrakech, Morocco.

Veronika LUX-POGODALLA and Alain POLGUÈRE (2011), Construction of a French lexical network: methodological issues, in *Proceedings of WoLeR 2011*, pp. 54–61, Ljubljana, Slovenia.

Katja MARKERT and Malvina NISSIM (2007), SemEval-2007 task 08: metonymy resolution at SemEval-2007, in *Proceedings of SemEval 2007*, pp. 36–41, Prague, Czech Republic.

Yann MATHET, Antoine WIDLÖCHER, and Jean-Philippe MÉTIVIER (2015), The unified and holistic method Gamma (γ) for inter-annotator agreement measure and alignment, *Computational Linguistics*, 41(3):437–479.

Igor MEL'ČUK (2010), La phraséologie en langue, en dictionnaire et en TALN, in *Proceedings of TALN 2010 (invited talks)*, Montréal, Canada.

Igor MEL'ČUK (2012), Phraseology in the language, in the dictionary, and in the computer, *Yearbook of Phraseology*, 3:31–56.

Marie MIKULOVÁ, Alevtina BÉMOVÁ, Jan HAJIČ, Eva HAJIČOVÁ, Jiří HAVELKA, Veronika KOLÁŘOVÁ, Lucie KUČOVÁ, Markéta LOPATKOVÁ, Petr PAJAS, Jarmila PANEVOVÁ, Magda RAZÍMOVÁ, Petr SGALL, Jan ŠTĚPÁNEK, Zdeňka UREŠOVÁ, Kateřina VESELÁ, and Zdeněk ŽABOKRTSKÝ (2006), Annotation on the tectogrammatical level in the Prague Dependency Treebank.

Annotation manual, Technical report 30, ÚFAL MFF UK, Prague, Czech Republic.

Joakim NIVRE, Marie-Catherine DE MARNEFFE, Filip GINTER, Yoav GOLDBERG, Jan HAJIČ, Christopher D. MANNING, Ryan MCDONALD, Slav PETROV, Sampo PYYSALO, Natalia SILVEIRA, Reut TSARFATY, and Daniel ZEMAN (2016), Universal Dependencies v1: a multilingual treebank collection, in *Proceedings of LREC 2016*, pp. 1659–1666, Portorož, Slovenia.

Aurélie NÉVÉOL, Cyril GROUIN, Jeremy LEIXA, Sophie ROSSET, and Pierre ZWEIGENBAUM (2014), The QUAERO French medical corpus: a resource for medical entity recognition and normalization, in *Proceedings of BioTxtM 2014*, pp. 24–30, Reykjavik, Iceland.

Marie-Sophie PAUSÉ (2017), *Structure lexico-sentaxique des locutions du français et incidence sur leur combinatoire*, Ph.D. thesis, Université de Lorraine, Nancy, France.

Alain POLGUÈRE (2014), Principes de modélisation systémique des réseaux lexicaux, in *Proceedings of TALN 2014 (volume 1: long papers)*, pp. 79–90, Marseille, France.

Carlos RAMISCH, Silvio Ricardo CORDEIRO, Agata SAVARY, Veronika VINCZE, Verginica BARBU MITETELU, Archana BHATIA, Maja BULJAN, Marie CANDITO, Polona GANTAR, Voula GIOULI, Tunga GÜNGÖR, Abdelati HAWWARI, Uxoá IÑURRIETA, Jolanta KOVALEVSKAITĖ, Simon KREK, Timm LICHTÉ, Chaya LIEBESKIND, Johanna MONTI, Carla PARRA ESCARTÍN, Behrang QASEMIZADEH, Renata RAMISCH, Nathan SCHNEIDER, Ivelina STOYANOVA, Ashwini VAIDYA, and Abigail WALSH (2018), Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions, in *Proceedings of LAW-MWE-CxG-2018*, pp. 222–240, Santa Fe, USA.

Carlos RAMISCH, Alexis NASR, André VALLI, and José DEULOFEU (2016), DeQue: a lexicon of complex prepositions and conjunctions in French, in *Proceedings of LREC 2016*, pp. 2293–2298, Portorož, Slovenia.

Victoria ROSÉN, Gyri Smørdal LOSNEGAARD, Koenraad DE SMEDT, Eduard BEJČEK, Agata SAVARY, Adam PRZEPIÓRKOWSKI, Petya OSENOVA, and Verginica BARBU MITETELU (2015), A survey of multiword expressions in treebanks, in *Proceedings of TLT 2015*, pp. 179–193, Warsaw, Poland.

Ivan A. SAG, Timothy BALDWIN, Francis BOND, Ann A. COPESTAKE, and Dan FLICKINGER (2002), Multiword expressions: a pain in the neck for NLP, in *Proceedings CICLing 2002*, pp. 1–15, Springer-Verlag, ISBN 3-540-43219-1.

Benoît SAGOT, Marion RICHARD, and Rosa STERN (2012), Annotation référentielle du corpus arboré de Paris 7 en entités nommées, in *Proceedings of JEP-TALN-RECITAL 2012 (volume 2)*, pp. 535–542, Grenoble, France.

Benoît SAGOT and Rosa STERN (2012), Aleda, a free large-scale entity database for French, in *Proceedings of LREC 2012*, pp. 1273–1276, Istanbul, Turkey.

Agata SAVARY, Marie CANDITO, Verginica Barbu MITITELU, Eduard BEJČEK, Fabienne CAP, Slavomír ČÉPLÖ, Silvio Ricardo CORDEIRO, Gülşen ERYİĞİT, Voula GIOULI, Maarten VAN GOMPEL, Yaakov HACHOHEN-KERNER, Jolanta KOVALEVSKAITĖ, Simon KREK, Chaya LIEBESKIND, Johanna MONTI, Carla Parra ESCARTÍN, Lonneke VAN DER PLAS, Behrang QASEMIZADEH, Carlos RAMISCH, Federico SANGATI, Ivelina STOYANOVA, and Veronika VINCZE (2018), PARSEME multilingual corpus of verbal multiword expressions, in Stella MARKANTONATOU, Carlos RAMISCH, Agata SAVARY, and Veronika VINCZE, editors, *Multiword expressions at length and in depth: extended papers from the MWE 2017 workshop*, pp. 87–147, Language Science Press, Berlin, Germany.

Agata SAVARY, Carlos RAMISCH, Silvio CORDEIRO, Federico SANGATI, Veronika VINCZE, Behrang QASEMIZADEH, Marie CANDITO, Fabienne CAP, Voula GIOULI, Ivelina STOYANOVA, and Antoine DOUCET (2017), The PARSEME shared task on automatic identification of verbal multiword expressions, in *Proceedings of MWE 2017*, pp. 31–47, Valencia, Spain.

Agata SAVARY, Jakub WASZCZUK, and Adam PRZEPIÓRKOWSKI (2010), Towards the annotation of named entities in the National Corpus of Polish, in *Proceedings of LREC 2010*, pp. 3622–3629, Valetta, Malta.

Nathan SCHNEIDER, Spencer ONUFFER, Nora KAZOUR, Nora EMILY DANCIK, Michael T. MORDOWANEC, Henrietta CONRAD, and Noah A. SMITH (2014), Comprehensive annotation of multiword expressions in a social web corpus, in *Proceedings of LREC 2014*, pp. 455–461, Reykjavik, Iceland.

Djamé SEDDAH, Reut TSARFATY, Sandra KÜBLER, Marie CANDITO, Jinho D. CHOI, Richárd FARKAS, Jennifer FOSTER, Iakes GOENAGA, Koldo Gojenola GALLETEBEITIA, Yoav GOLDBERG, Spence GREEN, Nizar HABASH, Marco KUHLMANN, Wolfgang MAIER, Joakim NIVRE, Adam PRZEPIÓRKOWSKI, Ryan ROTH, Wolfgang SEEKER, Yannick VERSLEY, Veronika VINCZE, Marcin WOLIŃSKI, Alina WRÓBLEWSKA, and Eric VILLEMONTÉ DE LA CLERGERIE (2013), Overview of the SPMRL 2013 shared task: a cross-framework evaluation of parsing morphologically rich languages, in *Proceedings of SPMRL 2013*, pp. 146–182, Seattle, USA,

Livnat Herzig SHEINFUX, Tali Arad GRESHLER, Nurit MELNIK, and Shuly WINTNER (2019), Verbal multiword expressions: idiomaticity and flexibility, in Yannick PARMENTIER and Jakub WASZCZUK, editors, *Representation and parsing of multiword expressions: current trends*, pp. 35–68, Language Science Press, Berlin, Germany.

Erik F. TJONG KIM SANG (2002), Introduction to the CoNLL-2002 shared task: language-independent named entity recognition, in *Proceedings of CoNLL 2002*, volume 20, pp. 1–4, Taipei, Taiwan.

Erik F. TJONG KIM SANG and Fien DE MEULDER (2003), Introduction to the CoNLL-2003 shared task: language-independent named entity recognition, in *Proceedings of CoNLL 2003*, pp. 142–147, Edmonton, Canada.

Agnès TUTIN and Emmanuelle ESPERANÇA-RODIER (2019), The difficult identification of multiworld expressions: from decision criteria to annotated corpora, in *Computational and corpus-based phraseology*, pp. 404–416, Springer-Verlag, ISBN 978-3-030-30135-4.

Agnès TUTIN, Emmanuelle ESPERANÇA-RODIER, Manolo IBORRA, and Justine REVERDY (2016), Annotation of multiword expressions in French, in *Proceedings of EUROPHRAS 2015*, pp. 60–67, Malaga, Spain.

Maarten VAN GOMPEL and Martin REYNAERT (2013), FoLiA: a practical XML format for linguistic annotation – a descriptive and comparative study, *Computational Linguistics in the Netherlands Journal*, 3:63–81.

Veronika VINCZE, István NAGY T., and Gábor BEREND (2011), Multiword expressions and named entities in the Wiki50 corpus, in *Proceedings of RANLP 2011*, pp. 289–295, Hissar, Bulgaria.

Marie Candito

© 0000-0001-8306-4859
marie.candito@u-paris.fr

Université de Paris, CNRS, LLF, Paris,
France

Mathieu Constant

© 0000-0002-9910-594X
Mathieu.Constant@
univ-lorraine.fr

Université de Lorraine, CNRS, ATILF,
Nancy, France

Carlos Ramisch

© 0000-0001-7466-9039
carlos.ramisch@lis-lab.fr

Aix Marseille Univ,
Université de Toulon,
CNRS, LIS, Marseille, France

Agata Savary

© 0000-0002-6473-6477
agata.savary@univ-tours.fr

Université de Tours, LIFAT, Tours,
France

Bruno Guillaume

① 0000-0001-8314-8075

Bruno.Guillaume@loria.fr

Université de Lorraine,
CNRS, Inria, LORIA, Nancy, France

Yannick Parmentier

① 0000-0003-1461-5535

yannick.parmentier@

univ-lorraine.fr

Université de Lorraine, CNRS, LORIA,
Nancy, France

Université d'Orléans, LIFO, Orléans,
France

Silvio Ricardo Cordeiro

① 0000-0002-1262-369X

silvioricardoc@gmail.com

Université de Paris, CNRS, LLF, Paris,
France

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier and Silvio Ricardo Cordeiro (2020), *A French corpus annotated for multiword expressions and named entities*, *Journal of Language Modelling*, 8(2):415–479

① <https://dx.doi.org/10.15398/jlm.v8i2.265>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

© <http://creativecommons.org/licenses/by/4.0/>