Catalan Journal of Linguistics Special Issue, 2019 **27-44**

Towards Matrix Syntax[*]

Roger Martin[†]
Yokohama National University
martin@ynu.ac.jp

Román Orús
Johannes Gutenberg-Universität
Donostia International Physics Center / Ikerbasque Foundation for Science
roman.orus@dipc.org

Juan Uriagereka
University of Maryland
juan@umd.edu

Received: December 31, 2017
Accepted: September 23, 2019

Abstract

Matrix syntax is a model of syntactic relations in language, which grew out of a desire to understand chains. The purpose of this paper is to explain its basic ideas to a linguistics audience, without entering into too many formal details (for which cf. Orús et al. 2017). The resulting mathematical structure resembles some aspects of quantum mechanics and is well-suited to describe linguistic chains. In particular, sentences are naturally modeled as vectors in a Hilbert space with a tensor product structure, built from 2x2 matrices belonging to some specific group. Curiously, the matrices the system employs are simple extensions of customary representations of the major parts of speech, as [±N, ±V] objects.

Keywords: syntax; chains; minimalist program; Hilbert space; matrix

[*] It is with a heavy heart that we send this paper to print, since our dear friend and co-author Roger Martin is no longer with us to grace us with his insight, rigor, and sheer decency. We have no doubt that he would have caught issues with the final version that are not apparent to us. We can at least assure the readers that Roger was with our effort till the very end, making it all worthwhile for us to simply pass the torch.

We thank Ángel Gallego and Dennis Ott for their efforts in editing this volume and organizing GYNSYN17, the conference from which it stems. We appreciate comments we received from participants, especially Tim Hunter, Marc Richards, and Masaya Yoshida. In addition, we are grateful to Sergio Balari, David Berlinski, John Colarusso, Ángel Gallego, Bill Idsardi, Michael Jarret, Diego Krivochen, Dan Lathrop, Steve Marcus, Dave Medeiros, Doug Saddy, and Zach Stone. Some of the research leading to this paper received financial support from a Grant-in-aid for Scientific Research from the Japan Society for the Promotion of Science (KAKEN 26370446; PI: Martin). Finally, our appreciation goes to an anonymous reviewer for insightful suggestions that helped us clarify parts of the paper.

ISSN 1695-6885 (in press); 2014-9718 (online) https://doi.org/10.5565/rev/catjl.221

Resum. *Cap a la sintaxi de matrius*

La sintaxi de matrius és un model formal de relacions sintàctiques en el llenguatge que va sorgir del desig de modelar les cadenes. L'objectiu d'aquest treball és explicar les idees bàsiques d'aquest model a un públic lingüístic, sense entrar en gaires detalls formals (vegeu Orús et al. 2017). L'estructura matemàtica resultant s'assembla a alguns aspectes de la mecànica quàntica i s'adapta bé per descriure les cadenes lingüístiques. En particular, les oracions es modelen naturalment com a vectors en un espai de Hilbert amb una estructura de producte tensorial, construïdes a partir de matrius 2 x 2 que pertanyen a un grup específic. Curiosament, les matrius que utilitza el sistema són extensions simples de representacions habituals de les parts principals del discurs com a objectes [± N, ± V].

Paraules clau: sintaxi; cadenes; programa minimalista; espai de Hilbert; matriu

Table of Contents

1. Preliminaries focusing on the Trouble with Chains
2. The Fundamental Assumption and Anti-Symmetrical Merge
3. Projecting from the Bottom and Selection Restrictions
4. The Explosion Problem with Specifiers and the Need for Matrix Compression
5. Chains and Beyond…
References

1. Preliminaries focusing on the Trouble with Chains

While it may not be necessary for an analytical science to be quantitative, gaining quantitative traction—if natural within a discipline's subject matter—can be an advantage. This is because of the rigor one can associate to calculations, but more generally because the level of predictions and accuracy of testing can gain a different scope. In practical terms, our project may be seen as a way to implement that desideratum within well-known presuppositions.

We assume the basic tenets of generative grammar, such as lexical categories, phrases, Merge, Agree, displacement, chains, control, ellipsis, rules of construal and other such notions that have arisen from a long tradition of theoretical investigations into the structure of the human language faculty. All of the familiar machinery that theoretical syntacticians commonly use constitutes our basic repository as well.

Within such a framework, one of the primary concerns is how to account for displacement phenomena, and other long-range correlations, particularly working within a so-called minimalist approach to grammar. Paramount among the issues is the fact that the interpretation of displaced objects is *distributed* (in phonetic and semantic terms). Although there has been much discussion about the reason for this over the years, little has been achieved in the way of understanding. In short, we know of no analysis that can account for the kinds of facts that we review below in classical computational terms (see Colins & Stabler 2016 for essentially the same admission).

Consider, for instance, the sentence involving multiple raising in (1), which can have either of the possible interpretations in A or B:

(1) Friends of each other seemed to the Obamas to appear to the Bushes to have shown up unannounced at the White House.

a. Friends of Barack seemed to Michelle and friends of Michelle seemed to Barack to appear to the Bushes to have shown up unannounced at the White House.

b. It seemed to the Obamas that friends of George appeared to Laura, and friends of Laura appeared to George to have shown up unannounced at the White House.

The availability of these two interpretations is commonly attributed to the assumption that displacement of syntactic phrases creates *copies*, as in (2):

(2) **Friends of each other** seemed to the Obamas **friends of each other** to appear to the Bushes **friends of each other** to have shown up unannounced.

If we focus on just one of the copies as the locus of interpretation, we can consider the three possibilities in (3), where here, to illustrate, the copy that is interpreted is highlighted in bold-face and the others are in strike-out.

(3) a. ~~Friends of each other seemed~~ to the Obamas **friends of each other** to appear to the Bushes ~~friends of each other~~ to have shown up unannounced.

b. ~~Friends of each other~~ seemed to the Obamas ~~friends of each other~~ to appear to the Bushes **friends of each other** to have shown up unannounced.

c. **Friends of each other** seemed to the Obamas ~~friends of each other~~ to appear to the Bushes ~~friends of each other~~ to have shown up unannounced.

If the interpretive component utilizes the bold-faced copy in (3a), this allows for binding of *each other* by *the Obamas*, yielding interpretation A. Whereas interpreting the bold-faced copy in (3b) allows *each other* to be bound by *the Bushes*, yielding interpretation B. Accessing the bold-faced copy in (3c) presumably does not yield any possible interpretation, assuming *each other* is not bound in that position.

However, many questions arise. First, as should be obvious, only one copy survives at the phonetic interface (PF), but the reason for this is unclear (why are not all copies pronounced/interpreted?). Furthermore, the copy that gets pronounced at PF necessarily corresponds to the bold-faced one in (3c), but we have seen how the bold-faced copies in (3a) or (3b) can also be accessed for interpretation at the semantic interface (LF). Yet, although all of the choices in (3) may be possible at LF, it seems that there too *only one* of the copies can be interpreted.[1] If that were

[1] It might seem that more than one copy is needed at LF when the copy used for scope/binding is not the same as the one involved in theta-role assignment. However, we assume theta-roles to be determined in a separate component of the grammar, as in Uriagereka (2008) and Martin and Uriagereka (2014). Hornstein (1998, 2001, etc.) advocates yet another approach to theta-roles that

not the case, arguably the unacceptable (4a) should be possible, with an LF like (4b).

(4) a. *Friends of each other seemed to themselves to have shown up unannounced.

b. $[_{IP}$ [friends of [each other]$_i$]$_i$ seemed to [themselves]$_i$ $[_{IP}$ [friends of [each other]$_i$]$_i$ to have shown up unannounced]]

With both copies of *friends of each other* available at LF, the higher copy could bind *themselves* and, at the same time, *each other* in the lower copy could be bound by *themselves*.[2]

Why it should be the case that no more than one copy is available for interpretation is no more obvious in the case of LF than is the question of why only one of the copies can be pronounced at PF. One can of course stipulate as much—but the question is why the objects behave *that way*, and not in other equally rational ways (all copies are interpreted, some copies are interpreted…). No formalism we know of yields that as a straightforward consequence.

We argue that chains (a set of occurrences of the same syntactic token created by a grammatical transformation) are *non-classical objects*, of the sort commonly assumed in physics, exhibiting conditions that have been described as "spooky" (see Martin & Uriagereka 2008, 2014 for the explicit statement of this idea). We are not the first to bring such notions into the discussion of language. For example, Paul Smolensky (Smolensky 1990; Smolensky & Legendre 2006) has argued for something along these lines for phonology and other parts of language—although within connectionist presuppositions that we do not find necessary. Researchers of various orientations have suggested similar "spooky" connections (e.g. Aerts & Aerts 1994; Atmanspacher et al. 2002; Bruza 2009; Coecke et al. 2013; Heunen et al. 2013; Gerth & beim Graben 2009; Khrennikov 2006; Piattelli-Palmarini & Vitiello 2015; Witteck et al. 2013, etc.).

More generally, we will be taking syntax to act on some Hilbert space, by way of linear operations. Moreover, projecting syntactic stuff into interface observables is what "collapses it" into a classical reality, in which entities present reference and quantification, truth values, or for that matter the very signals of speech (or writing, in most systems) are linearized one right after the other—which we also take to be a form of "collapse". That is our project in a nutshell.

None of this really makes sense without quantitative, or at least elaborate logical, assumptions. We think there is a simple way of proceeding, stemming from a

divorces them from intentional (scope/binding) semantics, treating them as features that get checked in the course of a derivation.

[2] The indexation in (4b) might be said to violate the so-called *i-within-i* prohibition. But the status of that condition, which was originally stipulated for theory internal reasons that are not obviously relevant today (and which furthermore incorrectly rules out grammatical expressions such as *Escher drew a picture of itself*), is far from clear. There are also other sorts of examples that, while being more complex for presentational purposes, demonstrate the same point and do not involve *i-within-i* situations. See, for example, Hornstein (1998).

fact that is familiar to linguists: *we operate on feature matrices.* We show below that it is easy to translate between familiar syntactic categories and matrices, and moreover that connatural to the latter, if their values are numerical, are interesting quantities that turn out to be central to a project attempting to construct relevant Hilbert spaces and, more generally, turn relevantly quantitative.

2. The Fundamental Assumption and Anti-Symmetrical Merge

Consider familiar objects as in (5), from Chomsky (1974).

(5) a. noun: [+N, -V] b. verb: [-N, +V]
   c. adjective: [+N, +V] d. adposition: [-N, -V]

(5) capitalizes on the intuition that "nouniness" is conceptually orthogonal to "verbiness", and those two separate lexical dimensions articulate all of the conceptual space the lexicon needs. N and V features were postulated by Chomsky so as to rationalize the distribution of lexical categories. He could, of course, have called those features A and B, or *1* and *i*, and retain the system we customarily teach our students. Note, however, that in the latter instance, i.e. *1* and *i*, there would be a greater level of precision: we could state the intuitive orthogonality of N and V in precise terms, inasmuch as *1* is mathematically orthogonal (maximally different) from $i = \sqrt{-1}$.

We make the Fundamental Assumption in (6) which leads to the reformulation of (5) as (7).

(6) *Fundamental Assumption:* N = *1* and $V = i = \sqrt{-1}$.

(7) a. noun: [1, -*i*] b. verb: [-1, *i*]
   c. adjective: [1, *i*] d. adposition: [-1, -*i*]

The representations in (7) may be seen as row vectors, in the sense that they are 1d arrays of numbers. While we could develop most of our formalism with such objects, for operational reasons we will "translate" the vectors in (7) to diagonal square matrices as in (8). (For an introduction to basic facts on linear algebra, see, e.g., <http://math.mit.edu/~gs/linearalgebra/>.)

(8) a. noun: $\begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix}$ b. verb: $\begin{pmatrix} -1 & 0 \\ 0 & i \end{pmatrix}$

   c. adjective: $\begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}$ d. adposition: $\begin{pmatrix} -1 & 0 \\ 0 & -i \end{pmatrix}$

Notice that the numerical values from the vectors in (7) are placed in the *matrix diagonal* in (8). As a result, (8) presents "diagonal" and "unitary" matrices, which means they have the property of their *inverse* being identical to what is called their *adjoint*. Explaining that technically would take us too far afield, but simply to fix some notation, for e.g. the noun matrix N = its *transpose* is N^T = (where we just exchange the antidiagonal elements, which in this case are both 0), and its *adjoint* (or *conjugate transpose*) is N+ = , where we just took the transpose matrix N^T and took the complex conjugate of its entries—i.e., replaced by Suffice it to say in this context that these are *extremely elegant matrices*, with well known mathematical properties. The formal objects in (8), which we call *Chomsky matrices*, in fact integrate into a curious mathematical group that will be discussed below. We may also add that these matrices form a non-standard basis for the Hilbert space C^2, which will also be discussed later.

To illustrate one of the merits of treating lexical categories in terms of the Chomsky matrices, consider the highly limited combinatorial possibilities for lexical heads and their complements:

(9) a. Nouns select PPs. b. Verbs select NPs.
c. Adjectives select PPs. d. Prepositions select NPs.

As generic statements, (9c) and (9d) are virtual universals, while (9a) and (9b) are at least statistically overwhelming. Certainly, verbs also select other categories, which may force us to complicate the system—and also to invoke functional categories—but in science one typically begins by trying to predict the most basic interactions.[3] In any event, while the facts in (9) are commonly stipulated in many ways, we have never seen them explained.

To provide an account for (9), we begin with the assumption that First-Merge—the combination of a head and its complement—is *matrix multiplication*. Indeed, once we postulate the Fundamental Assumption in (6), and represent categories as in (8), it is natural to ask whether those moves lead to more than just the formalization of orthogonality of N and V features, or standard operations among them are also possible. Multiplication is, in a sense, a deformation of a given (conceptual) space by way of a linear operator. One should then ask why matrix multiplication should model First-Merge (or any other process). While there is no *a priori* answer to such a question, we can attempt to show the predictive results of taking the step.

We further assume the following:

(10) First-Merge is antisymmetrical.

Typically, First-Merge creates an *asymmetrical* relation, in the sense that one element, the head, is necessarily *atomic* (selected from the lexicon), whereas the other, the complement, is a *complex* phrasal element that has been previously

[3] It could also be that (9) is in some sense *cognitively prior*, in that learners get it from Universal Grammar in the absence of experience. From that perspective, further complications would be learnt or emerge only later, based on more complex interactions with environmental stimuli.

assembled in the derivation. However, there is one notable exception to this situation, corresponding to the initial combinatorial step of every derivation (or the initial step in the sub-derivation of, say, a left branch constituent, etc.). Obviously, at the very start of a derivation there cannot be any previously assembled syntactic objects. Thus, the only option then is to combine two lexical items. The idea behind (10) is that we can allow for this sort of situation—while still maintaining the general asymmetry of the head-complement relation—*if we restrict the initial combination of two lexical items to instances of self-merge*, an idea first proposed by Guimarães (2000) and later adopted by Kayne (2009).[4]

When considered from the point of view of the Chomsky matrices in (8) and first-merge as matrix multiplication, the result of self-merging any of the Chomsky categories is the same:[5]

(11)
$$\begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & i \end{pmatrix} \cdot \begin{pmatrix} -1 & 0 \\ 0 & i \end{pmatrix} =$$
$$= \begin{pmatrix} -1 & 0 \\ 0 & -i \end{pmatrix} \cdot \begin{pmatrix} -1 & 0 \\ 0 & -i \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = Z$$

Z is one of the famous Pauli matrices, which has been put to use to predict properties of an electron's angular momentum. The reasons for that are not important now, but they boil down to the fact that *Z* is what mathematicians call a *Hermitian matrix*.

Hermitian matrices are to matrices what real numbers are to numbers, in that eigenvalues (roots) of Hermitian matrices are real. Both can be measured. We will get a feel for that as we get our hands into computations, but we can point out the obvious already: the elements in the diagonal in *Z* are both real numbers. These are key in understanding the essence of a matrix: its eigenvalues. The eigenvalues of the Chomsky matrices are combinations of ±1 and ±*i*. It is different for *Z*, as a result of which the matrix has other elegant properties. We can think of *Z* as a welcome encounter arising from the self-merger of the Chomsky objects. At the same time, it is also interesting to ponder what we should make of that, especially within an ultimately "semiotic" system that in some sense *carries thought*, and even allows us to communicate it.

Basically, a linguistic system that is trying to start in a self-merger with the math in (11) has to resolve that "multiguity", so that instead of all possible self-mergers leading to *Z*, the system chooses one—*any* one—to the exclusion of the others. One may think of this choice as the core *Saussurean arbitrariness* in the system, as the

[4] A relation that is asymmetrical except when holding with itself is called *antisymmetrical*. Here we state that First-Merge is antisymmetric, though precisely speaking the property holds of *the relation* established between the two elements that undergo First-Merge (not the assembling operation).

[5] To multiply such matrices as in (11), readers may multiply entries in entry-wise fashion. But note that this is possible only because the matrices are diagonal (it is not meant as a general comment about matrix products).

choice of any such mapping is in principle as good as any other. The choice made by language seems to be the following:

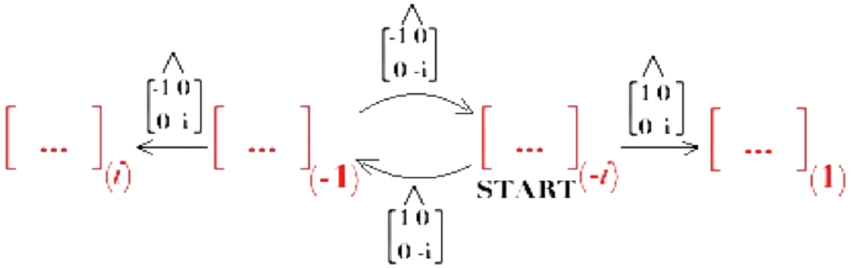
(12) N (understood as Chomsky's $\begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix}$) self-merges as $\begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix}$ $= Z.$

(12) can be viewed as *cognitive anchor*, which we will not seek to explain, at least not mathematically. We assume (12), as opposed to any of the other logically possible mappings, for empirical reasons: derivations bottom out as nouns.[6] That Guimarães proposed self-merger *for nouns* is not surprising; the insight was in the self-merger—not its involving nouns.

3. Projecting from the Bottom and Selection Restrictions

Once the "anchor" for human language in (12) is assumed, things start falling into place, in a form that can be summarized in terms of a diagram proposed to us by Michael Jarret, which we refer to as the *Jarret graph*, presented abstractly in (13). In this graph, we need to distinguish operational ***edges*** (Chomsky matrices corresponding to the four major lexical heads, which are presented in the graph with a "hat" ^ to signal their operator status) and ***nodes***. Both edges and nodes are matrices, but the emphasis in each instance is entirely different: while the edges are unique linear operators, hence they are presented with fully specified values, the matrices they operate on as vectors can be of any kind that presents, in matrix representation, a given *determinant*, signaled in parenthesis, ranging over ±1 and ±*i*.

(13)



A matrix determinant is an invariant scalar obtained, for simple square matrices, by multiplying the items in the diagonal and subtracting from that the product of the items in the off diagonal. We propose, for matrices understood as vectors, that the matrix determinant determines what linguists call the "label" of a category, which for the projections we will be operating with are *the fundamental orthogonal*

[6] Also, of the lexical categories, *only nouns appear in bare form*, without dependents, as names, pronouns, etc.

features ±1 and ±i only. Importantly, we assume that the interpretation of determinants as labels is only relevant for the outputs of operations (namely, the nodes in the graph, which correspond to phrasal projections consisting of a lexical head and complement, except for the initial step of self-merge), and not the operators (or lexical heads) themselves.[7] The specific labeling system we argue for projected categories is the following:

(14) a. N projections: label/determinant -1 b. V projections: label/determinant *i*

c. A projections: label/determinant 1 d. P projections: label/determinant -*i*

The Jarret graph has N heads select (multiply with) matrices of type -*i* (prepositional projections) to yield -1 projections, while P heads select matrices of type -1 (the nominal projections) to yield -*i* projections—that being the *recursive core* of the system. In addition, the graph also says that V heads select matrices of type -1 to yield *i* projections, while A heads select matrices of type -*i* to yield 1 projections—that being the *non-recursive periphery* of the first-merge system.

In addition, the graph has a START point, explicitly signaled in (13). This boils down to the *anchoring assumption* we have argued for. It would be silly for a graph as in (13) to start at the periphery, since then the computation has nowhere to go; the core is a more useful place to start. But the core itself has *two* different sites: one labeled -*i* and the other one labeled *1*. On formal grounds alone, it is natural for the system to start at a state that *carries the computation to the very elegant Pauli matrix Z*, with determinant/label -*1*. We have already shown above how all instances of self-merge, for any of the Chomsky categories, yield this result. That being the case, the *only* matrix that carries the system to the *Z* configuration with determinant -1 is precisely $\begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix}$, which we call Chomsky1 (or C1), for this very reason. (It is still a substantive claim to postulate that C1 corresponds specifically *to nouns*, which we are adapting from Chomsky 1974 via our Fundamental Assumption; the formal system could have just as naturally started in C1 with us having assigned that matrix to verbs, prepositions or adjectives…)

The other formal properties of the Jarret graph—such as why the -*1* and -*i* projections are at the core, others at the periphery—*follow from the results of matrix multiplication over the Chomsky matrices.* Specifically, only the following *eight* results are mathematically possible, via multiplication. We have mentioned how, starting on the self-merge of Chomsky's *C1* (15e), we obtain Pauli's *Z* (15a), with label/determinant -1. We can then proceed with the specific options in the Jarret graph. *Z* can multiply into -*C1* (15g) with label/determinant -*i* by -*C2* (15h) (stay-

[7] We emphasize again that the matrices understood as linear operators are as distinct from what they operate on as, say, operator "+" is from the number pair it takes – "+" is not, itself, even a number, let alone a pair. In that representation, the label for our matrices-as-operators is *not* its determinant. In contrast, when a given matrix is interpreted as a vector that linear operators operate on, then its determinant is the object's label.

ing at the core of the graph), *or* multiply into -*C2* (15h) with label/determinant *i* by -*C1* (15g) (going into the left periphery of the graph).

(15) a. *Z* b. *I* c. -*Z* d. -*I* e. *C1* f. *C2*

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix},$$

g. -*C1* h. -*C2*

$$\begin{pmatrix} -1 & 0 \\ 0 & i \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -i \end{pmatrix}.$$

The same reasoning obtains for the ensuing matrices. For example, the -*C1* obtained in the previous instance can multiply into -*Z* (15c) with label/determinant -1 by *C1* (15e) (staying at the core of the graph), or it can multiply into -*I* (19d) with label/determinant 1 by *C2* (15f) (going into the right periphery of the graph). Readers can try this as an exercise for other states in the graph, and it will become apparent that *all the results fall within the "first-merge" Abelian group* in (15), which is commutative for multiplication.

Note that, for each of the lexical projections—matrices that the system treats as *vector arguments* of other matrix operators—there are *two equivalent matrix variants* with the same label/determinant, which we refer to as "twin" projections. The matrices-as-vectors are equivalent in that *they share the same determinant*, which we have proposed can be understood syntactically as a label. For example, *Z* and -*Z* have label/determinant -*1* because the determinant is the product of the items in the diagonal minus the product of those in the off diagonal—so -*1* in both instances. Readers can easily verify that this is true for all other twin categories in (16).

(16) a. NP, label -1 b. AP, label 1 c. VP, label *i* d. PP, label -*i*

Z -*Z* *I* -*I* *C2* -*C2* *C1* -*C1*

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -i \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & i \end{pmatrix}.$$

Other matrix multiplications are possible among the eight items in (16); but *only those expressed in the Jarret graph present this kind of symmetry*. For example, we could have multiplied, say, a preposition understood as a -*C1* operator (15g) times a verb phrase understood as matrix-as-vector *C2* (15f), one of the VP "twins"; the result is -*I* (15d) with label/determinant *1*; but while that is mathematically fine, it simply cannot be a projection from a prepositional head -*C2* with label -*i*. Readers can try similar multiplications off the edges of the Jarret graph, to see how only the connections made explicit within it preserve endocentricity/selection in the sense described.

That is what predicts the facts in (9), together with the formal fact that multiplication only allows certain results. Had we asked whether we could obtain a projected *Z* (16a) from the last matrix multiplication mentioned in the previous paragraph (*-C1* (15g) times *C2* (15f)), the answer would be *no*. That is not for substantive reasons as presupposed in (14); it follows from assuming a numerical base and elementary multiplications—one cannot obtain *i* from *1x1*. Thus there is an important consequence of the numerical assumptions we made to substantiate Chomsky's intuition about the cognitive orthogonality of N and V attributes, as well as the general approach to treating categories as feature matrices (in linear-operator interpretation for heads and vector interpretation with "twin" variants for projections), together with interpreting these and their hypothesized elements in a mathematical sense: we are now able to *predict certain elementary combinations in syntax without having to invoke external considerations.*

4. The Explosion Problem with Specifiers and the Need for Matrix Compression

Just as we have proposed matrix multiplication for first-merge, we now propose another kind of product, the *tensor product*, for Elsewhere Merge. We refer that way to those forms of Merge in which *both* of the merged elements are complex, having a derivational history (instead of one of them, at least, coming from the lexicon). The rhetoric in the literature routinely equates these two forms of merge, but we will keep them separate, to begin with because they are plainly distinct in that First Merge must involve a projecting head, whereas Elsewhere Merge does not. There are also many empirical differences between firstly merged complements and specifiers merged in elsewhere fashion—but we will not review them now. In any case, just as there is no *a priori* reason to treat a given form of merge as a type of product, there isn't one to treat another as a different type of product. The argument is ultimately based on how well the decision may lead to modeling relevant facts. In this regard, there are two broad themes to keep in mind from the perspective of modeling chains, which is our driving force. First, that a (displacement) chain by definition must involve at least one specifier—in that the displaced site cannot involve a form of First Merge. Second, that the payoff of the sort of math we are pursuing arises when considering tensorized networks, for formal reasons we review shortly.

Tensor products have the effect of basically *concatenating two matrices* into a larger one, which is useful in "building structure". Whereas regular matrix multiplications do not preserve structure (once modified, a linearly altered structure could have come from different multiplications), *tensor products are structure-preserving*: by looking at a tensor product, we know what went into it. For this reason, while matrix multiplication retains the dimensionality of its factors, tensor products generally have *a dimensionality that grows upon taking place,* as function of the dimensionality of the factors. The dimensionality of a matrix is its number of rows and columns—or the information that takes to specify it—determining what sorts of operations are allowed.[8] For the objects in the Abelian group in (15),

[8] For example, only matrices of identical inner dimensionalities can be added/subtracted, and only a matrix A with the same number of rows as the number of columns as a matrix B can enter into a multiplication A B.

multiplying its members times any other does not change the dimensionality of the factors: the result is of the same dimensionality of each of the factors—otherwise (16) would not be a group. But this does *not* happen when we invoke a tensor product concatenating two matrices. The dimensionality of the resulting matrix is *the product of the dimensionalities of the factors.*

(17) $$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 3 & 4 \end{pmatrix}$$

Had (17) been matrix multiplication involving the identity matrix *I*, the result would be identical (in dimensionality and everything else). Because this is a tensor product, even if it involves the identity matrix, the result preserves the *shape* of the second factor, but it is obvious that the output is a 4x4 matrix. Moreover, by looking at the output we know that it must have originated in the product to the left, in this sense the tensor product being structure preserving.

We use structure-preserving tensor products to generate phrase-to-phrase mergers (as opposed to head-to-phrase conditions). This has vast consequences. We generate (18a) by the tensor product of *children's* and *pictures of NYC*. This is possible regardless of whether the genitive is complex, as in *relatives of children's* (the situation in (18b))—a phrase like that falling into the characterization provided for (18a). But the situation in (18c) is more interesting:

(18) a. *Children's pictures of NYC.*
b. *Relatives of children's pictures of NYC.*
c. *Women's children's pictures of NYC.*
d. *London's women's children's pictures of NYC.*

In (18c) we have a specifier (*women's*) within another (*women's children's*). So if each elsewhere merger, going beyond the initial head-complement relations, invokes a tensor product, *and* a tensor product's dimensionality is the product of its factor's dimensionalities, then the dimensionality of *women's children's pictures…* should be equal to that of *children's pictures* **times** that of *women's*. This can then go on into the even higher dimensionality of *London's women's children's pictures…* as in (18d) and so on—indefinitely. We call this the Explosion Problem, which tells us something about the nature of specifiers. The general approach to such problems is matrix compression, based on *dimensional reduction*. What we seek for that purpose are matrix results where *entire rows or entire columns reduce to zero*, and thus can be eliminated. These, it turns out, correspond to matrices with one or more zero eigenvalues.

Consider next some generalities to be drawn about our "Magnificent Eight" objects in (16) (the Chomsky matrices, the Pauli matrix ±*Z*, the identity matrix and its inverse ±*I*) in particular:

Table 1. Algebraic properties of the Pauli matrices within the Magnificent Eight

Matrices: Properties	$Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$	$-Z = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$	$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$-I = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$
Char. polynomial	$x^2 – 1$	$x^2 – 1$	$x^2 – 2x + 1$	$x^2 + 2x + 1$
Eigenvalues	1, -1	-1, 1	1, 1	-1, -1
Determinant	-1	-1	1	1
Trace	0	0	2	-2

Table 2. Algebraic properties of the Chomsky matrices within the Magnificent Eight

Matrices: Properties	$C1 = \begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix}$	$-C1 = \begin{pmatrix} -1 & 0 \\ 0 & i \end{pmatrix}$	$C2 = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}$	$-C2 = \begin{pmatrix} -1 & 0 \\ 0 & -i \end{pmatrix}$
Char. polynomial	$x^2 – (1– i)x –i$	x^2 - $(-1+ i)x – i$	$x^2 – (1+ i)x + i$	x^2 - $(-1- i)x + i$
Eigenvalues	1, -*i*	-1, *i*	1, *i*	-1, -*i*
Determinant	-*i*	-*i*	*i*	*I*
Trace	1– *i*	-1+ *i*	1+*i*	-1- *i*

For these matrices, the following statement is always formally true:

(19) *The diagonal elements are the polynomial roots and matrix eigenvalues*.

Note also that the Pauli matrices (Table 1) are different from the Chomsky matrices (Table 2) in that *all their eigenvalues are real*. Hermitian matrices have real eigenvalues, so we can easily see that none of the Chomsky matrices are Hermitian. Observe, also, putative unifications across categories. We have already observed how the positive and negative versions of the "twin" categories, which we use in vector interpretation for category projections, share the same determinant. But there are more generalizations of interest. Note that only ±Z presents the *same characteristic polynomial* $x^2 – 1$ (all other matrices in the Magnificent Eight have different characteristic polynomials). Now that is a specific sense in which ±Z is the most elegant among the Magnificent Eight: aside from being Hermitian, it has a unified characteristic polynomial, a unified trace,[9] and a unified determinant—which no other matrix pair in the group does.

Needless to say, considerations about characteristic polynomials, eigenvalues, and so on, obtain for all square matrices, not just our Magnificent Eight. This is important when considering these architectural issues from a broader perspective,

[9] The notion *matrix trace* in these tables—which of course has nothing to do with syntactic traces—is just the *sum of the elements in the matrix diagonal*. To avoid a confusing notation we signal as *tr* the matrix trace.

including an extension of our system to functional categories. To begin with, Pauli's ±Z is only one of three Hermitian matrices within the Pauli Group, which includes also ±I and imaginary versions of all these matrices. Two other fundamental Pauli matrices, $\pm X = \pm \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\pm Y = \pm \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$, are non-diagonal. It turns out that when we multiply any of the non-diagonal matrices in the Pauli Group by any of the Chomsky matrices, we end up in another group of 32 matrices that we call *the Pauli/Chomsky Group*. This group is extremely interesting in its own right, and furthermore it allows us to systematically construct and explore a corresponding vector space where tensor products (specifier relations) can be superposed (forming chains). Moreover, the larger group allows us to postulate a richer *"periodic table" of syntactic elements*: the Magnificent Eight (lexical projections) along with *corresponding functional projections*, which also have well-defined characteristics of the sort just studied (some will have unified polynomials, not all; some will be Hermitian, not all; some will be unitary, etc.—all of which has consequences for a generalized semantic anchoring).

Within those parameters, we take the research program to be as follows:

(20) A. To find out how the functional categories (Infl, Comp, *v*, etc.) relate to lexical categories and to one another in a principled fashion.

B. To determine how *Pauli/Chomsky Group* and its 16 twin projections constitute the basis for standard syntax, in terms of their multiplications and tensor products.

C. To understand which of the tensor products among the categories in the periodic table lead to compressible results.

D. To figure out how the tensor products sum with one another into chain dependencies, and which among those present separable contexts.

Just as the formal tools in Tables 1 and 2 above show us in what sense the Pauli/Chomsky matrices are more or less elegant—which arguably relates to their syntactic distribution—they also allow us to understand how dimensions can be reduced after they have grown due to a tensor product. In this regard, it is useful to emphasize that our matrices can be said to have a dimensionality *equal to their number of non-zero eigenvalues.* In other words, a non-compressible 4x4 matrix has four substantive (not zero) eigenvalues, whereas a compressible 4x4 matrix has as many zero eigenvalues as matrix dimensions are irrelevant to it.

It will not be possible for us to present, in this context, anything but a "teaser" of how the dimensional reduction works—and we refer interested readers to Orús et al. (2017) for relevant details. The basic idea, however, works as follows. We can find situations in which literally *adding* a tensor product (arising from a projection taking a specifier) to another such product results in the elimination of some of the eigenvalues in the sum of the matrices. For example, with the sum shown in

(21), the specifications of which are given in (22), we clearly have a dimensional reduction, since *the ensuing matrix has two zero eigenvalues.*[10]

(21)
$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & i & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -i \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -i & 0 \\ 0 & 1 & 0 & 0 \\ i & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & i & -i & 0 \\ 0 & 1 & -1 & 0 \\ i & 0 & 0 & -i \end{pmatrix}$$

(22) dt.: *0,* tr.: *0;* char. pol.: $x^4 + 2i\,x^2$; eigenvalues: (-1+ i), (1 - i), 0, 0.

So the system will basically reduce the specifier dimensionality by integrating it into a sum of the sort in (21)—which we take to be a chain if certain structural conditions are met. Another way of putting this more intuitively is that the general rationale behind chains, from the present perspective, is to reduce specifier dimensionality.

5. Chains and Beyond…

The importance of the foregoing exercise is to prepare the ground for the operations that, in conditions of superposition (sums) as in (21)/(22), may lead not just to dimensionality reduction in specifiers—but also to different chain collapses. This is the crux of the idea: chains exist, prior to being observed, in *superposed states.* At the observation point, if at all possible, they materialize, with some probability, in one of those states, which thus becomes observable.

There are well understood properties of superposed states that, in principle, allow for their separability; for instance when they are *orthogonal* to start with (with regards to some orthonormal basis). The situation is all or nothing: if the states are orthogonal, the separation, in the right conditions, is inevitable; if they are not orthogonal, the separation is impossible. Moreover, there is no such thing as being observable in multiple states *at the same time*, much as there is meaning to the states all existing simultaneously.

With Chomsky (1995), we take a chain to be an object of the sort {{α, K}, {α, Λ}}, where a specifier α moves from context Λ to context K. Since we are modeling specifiers with tensor products, we can take the chain to be the sum:

(23) $[\alpha \otimes K] + [\alpha \otimes \Lambda] = \alpha \otimes [K + \Lambda]$

To say α *separates* from these superpositions is to say one can "factor out specifier α" from the relevant tensor products, as in the right-hand side of the equations in (24). So the chain, in a deep sense, *links the contexts of each of its occurrences*, K and Λ. After "factoring out" the separable element α, what remains is the super-

[10] For diagonal square matrices, the determinant amounts to the product of the eigenvalues, while the trace equals the sum of the eigenvalues. The matrix determinant and trace of the sum in (21) is zero, as a consequence of which the characteristic polynomial is simplified to $x^4 + 2i\,x^2$.

position [K + Λ]. Now here is the key: if the superposed contexts are mutually *orthogonal*, we can apply to such *complementary* conditions the standard logic in quantum mechanics. Basically, when the relevant system is measured, it has 50% probability of being observed in the K configuration and 50% probability of being observed in the Λ configuration. If we suppose that we observe abstract linguistic representations by sending them to relevant interfaces, *within those representations* we can say that chain {{α, K}, {α, Λ}} *collapses* at either configuration K or configuration Λ, with equal probability.

Of course, we have to make precise what we mean by "orthogonal", or "maximally different" within an orthonormal basis. The following is the standard approach:

(24) Two vectors *x* and *y* in vector space *V* are orthogonal if their inner (scalar) product is zero.

A convenient way to define the scalar product between two matrices is as in (25), where *tr* represents a matrix trace (and see fn. 9):

(25) $<A|B> = \mathrm{tr}(A^{\dagger} B)$
Where, for *ket* |A>, A's conjugate adjoint $A^{\dagger}$ is the *bra* <A|.

Here we are using a vector notation introduced by Paul Dirac for notions discussed above already. What (24)/(25) boil down to is that we take two matrices *A* and *B*, understood as vectors, to be orthogonal if and only if the trace of multiplying *A's adjoint A†* times *B* is zero. Because we have the Pauli/Chomsky group to work with, determining this, which in Dirac's shorthand is <A|B>, is relatively simple: we just need to churn the calculations.

The points to take home are straightforward. First, this is supposed to work with the very same types of conditions and reasoning as it does in quantum physics. The issue is not really whether the computations are wrong (they aren't), but rather whether they are meaningful. To decide on that depends on whether we have alternative theories of "reconstruction effects" and the like, and if so, whether such alternatives fare better on empirical grounds. Our attempt here is simply to show how things work in our terms.

Second, the formalism, as such, allows for little or no wiggle room. If "collapses" are meant seriously, they take place in a vector (Hilbert) space along the lines of what is guaranteed by (24)/(25) in the context of something like the Chomsky/Pauli group. In particular, if two matrices come out as orthogonal by the definitions we are introducing, they cannot be "quasi-orthogonal" or "orthogonal up to speaker intuitions", or whatever. One could, of course, change the definition of the inner/scalar product in (25), and then different things would be orthogonal. Or reject the Pauli/Chomsky group as the locus for all of this, and then perhaps in a different realm other things would be orthogonal. But within the scenario we are presenting there are *no* alternatives. A third broad point to bear in mind is that we are attempting several things at the same time. At the very least we need to address the Compression Problem for specifiers. This is to say that we are not just after "reconstruction effects" for chain occurrences. While that is what has motivated the program, once we invoke matrices, groups, Hilbert spaces and so forth, one hopes that all of that doesn't amount to mere paraphernalia to address the technical problem of occurrences. For us, chain occurrences are interesting inasmuch as they touch on all these other issues, taking us from humble phrases to complex long-range correlations.

To be sure, chains are not the only long-range correlations that grammars present: there is obligatory and non-obligatory control, ellipsis of various kinds, binding and obviation effects, preferences, and much more. We have the sense that treating these matters within a Hilbert space of relations is promising, particularly when, beyond the superpositions just discussed, such system a fortiori involves rampant entanglements. Basically, *whatever doesn't separate is entangled*, so there is plenty of room to explore what happens beyond the core situation discussed above. Present space limitations aside, the issue of entanglements is one that we are currently working on and do not have a full understanding of yet, to be honest.

One last point is worth emphasizing: much of what we have said above would not make (non-metaphorical) sense without the use we have made of scalars of different kinds. We have shown the role played by determinant scalars. We have just alluded to the important role of matrix traces—another scalar—in determining the inner product of our Hilbert space. (We could also show how traces in the Pauli/Chomsky matrices help us separate substantive categories from grammatical ones.) Moreover, the logic of chain collapses as sketched ultimately follows the usual logic that is also applied to quantum mechanics. That very logic requires a "lower boundary", usually expressed in terms of Planck's famous constant for the case of quantum physics—at any rate, *a non-zero real number*. That apparatus has to be numerical, indeed real in the technical sense. No real numbers, no syntax as we have examined it. We could certainly be wrong in our analyses, but if we are not, they provide *bona-fide* arguments that "mind phenomena" require real quantities as they materialize, enough at least to show up with coherent patterns as examined here.

As we noted already above, we are not the first to have argued that the human language utilizes a Hilbert space (or some extension) of some sort, or that it is best to treat some aspects of grammar in terms of vector spaces more generally. We believe, however, that we are the first to make such a "quantum leap" taking totally seriously the fundamentals of linguistic theory (the division into nouns, verbs, adjectives and adpositions, the role of structure, selection and endocentricity, within phrases, standard cartographies, etc.). This is a sense in which our approach is actually as conservative as it is admittedly radical. We have shown how a Hilbert space can be constructed from assumptions that many linguists teach in their undergraduate classes. The only twist we have added is to interpret familiar conceptual orthogonalities in mathematical terms, which we have found worth studying.

References

Aerts, D. & Aerts, S. 1994. Applications of quantum statistics in psychological studies of decision processes. *Foundations of Science* 1: 85-97.

Atmanspacher, H., Römer, H. & Walach, H. (2002). Weak quantum theory: Complementarity and entanglement in physics and beyond. *Foundations of Physics* 32: 379-406.

Bruza, P., K. Kitto, D. Nelson & C. McEvoy. 2009. Is there something quantum-like about the human mental lexicon? *Journal of Mathematical Psychology* 53: 363-377.

Collins, C. & E. Stabler. 2016. A Formalization of Minimalist Syntax. *Syntax* 19(1): 43-78.

Chomsky, N. 1974. *The Amherst Lectures*, delivered at the 1974 Linguistic Institute, University of Massachusetts, Amherst: Université de Paris VII.

Chomsky, N. 1995. *The Minimalist Program*. Cambridge: MIT Press.

Gerth, S. & P. beim Graben. 2009. Unifying syntactic theory and sentence processing difficulty through a connectionist minimalist parser. *Cognitive Neurodynamics* 3: 297-316.

Guimarães, M. 2000. In Defense of Vacuous Projections in Bare Phrase Structure. In Guimarães, M., L. Meroni, C. Rodrigues & I. San Martin (eds.). *University of Maryland Working Papers in Linguistics* 9: 90-115.

Heunen, C., M. Sadrzadeh & E. Grefenstette (eds.). 2013. *Quantum Physics and Linguistics*. Oxford: Oxford University Press.

Hornstein, N. 1998. Movement and Chains. *Syntax* 1(2): 99-127.

Hornstein, N. 2001. *Move! A Minimalist Theory of Construal*. Oxford: Blackwell.

Kayne, R. 2009. Antisymmetry and the Lexicon. *Linguistic Variation Yearbook 2008*: 1-32.

Khrennikov, A. 2006. Quantum-like brain: 'Interference of minds'. *Biosystems* 84(3): 225-241.

Martin, R. & J. Uriagereka. 2008. Competence for preferences. In X. Artiagoitia & J. A. Lakarra (eds.). *Festschrift for Patxi Goenaga.* University of the Basque Country.

Martin, R. & J. Uriagereka. 2014. Chains in Minimalism. In P. Kosta, S. Franks, T. Radeva-Bork & L. Schürcks (eds.). *Minimalism and Beyond: Radicalizing the Interfaces.* Amsterdam: John Benjamins.

Orús, R. & Martin, R. & Uriagereka, J. 2017. *Mathematical foundations of matrix syntax*. Retrieved from <https://arxiv.org/abs/1710.00372>.

Smolensky, P. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist networks. *Artificial Intelligence* 46: 159-216.

Smolensky, P. & G. Legendre. 2006. *The harmonic mind: From neural computation to Optimality-Theoretic grammar* (vols. 1-2). Cambridge: MIT Press.

Uriagereka, J. 2008. *Syntactic Anchors: On Semantic Structuring.* Cambridge: Cambridge University Press.