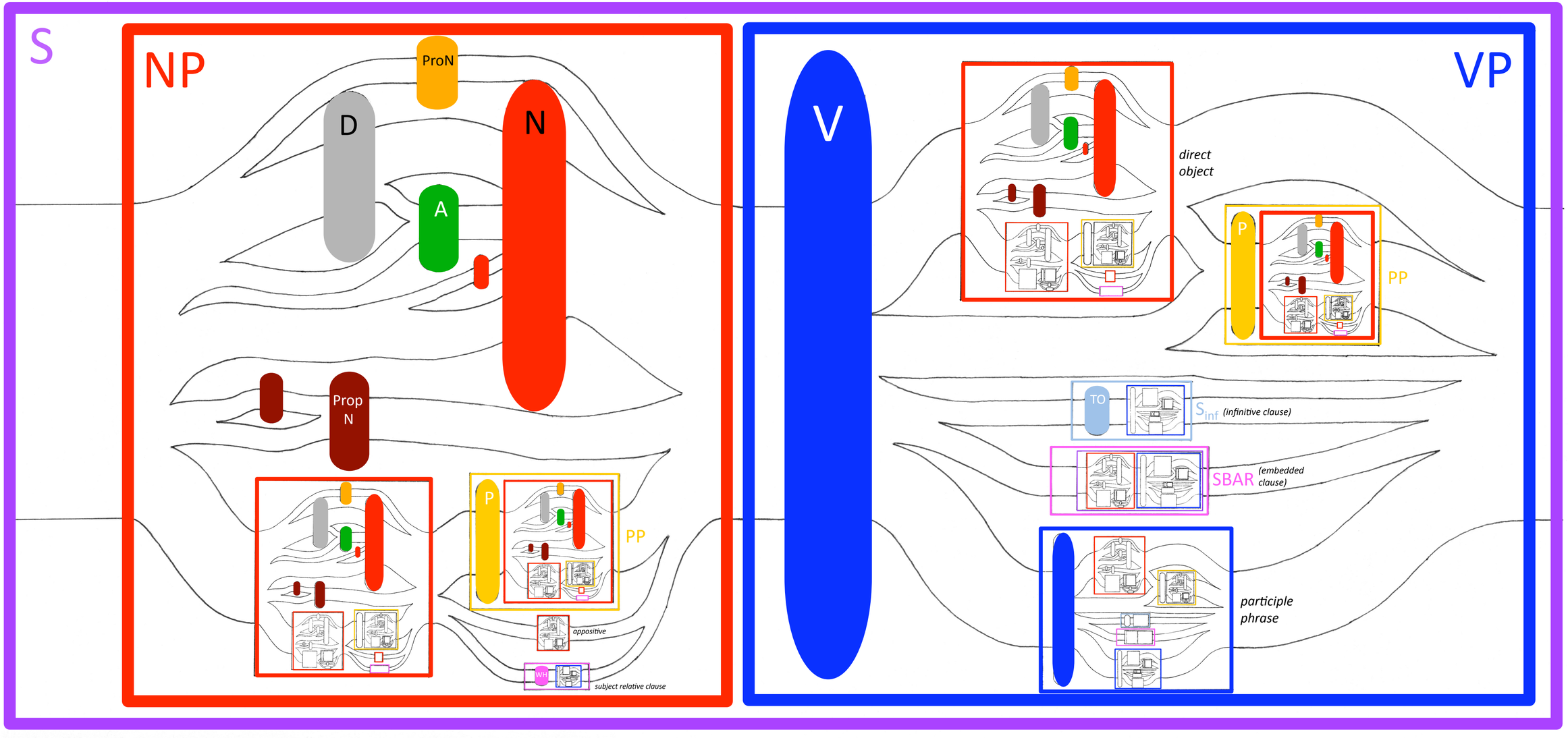
A VISUAL REPRESENTATION OF THE ENGLISH LANGUAGE.

By David Temperley



This poster presents a graphic representation of syntactic probabilities in written English. The probabilities were extracted from the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993)—about 1 million words of syntactically-annotated text. (The corpus was slightly modified for this purpose.)

The corpus can be used to define a *probabilistic context-free grammar*, in which each expansion of a constituent is assigned a probability. The diagram represents only the 25 most frequently-occurring expansions; these account for about 65% of all constituent tokens in the corpus (see Table 1 at right).

The diagram consists of a network of paths, boxes, and ovals. Paths represent syntactic expansions; boxes represent constituents; ovals represent preterminals (noun, verb, etc.). Each box has a path entering its left side and exiting its right side (consider the large NP on the left as an example). The path may split into several sub-paths, representing possible expansions. If the height of the path (i.e. the distance across in the vertical dimension as it enters the box) is *H*, the height of each sub-path is *P* × *H*, where *P* is the probability of the expansion. For example, the probability of the expansion NP → ProN (Pronoun) is .092, so the height of the ProN sub-path is .092 × the height of the NP path. Smaller ovals and boxes represent the "children" to which the larger constituent expands. (The height of each oval and box is 2.14 x the height of the path through it.) When multiple expansions involve the same child, their paths are merged into a single path whose height is the sum of the sub-paths; for example, the NP expansions [D N] and [D A N] both involve D, so their paths (going into D) are merged. Each smaller constituent also shows its internal structure; the probabilities of elements within it are correctly represented with respect to the immediate parent and higher-level constituents as well.

The diagram represents many important facts about English. It shows, for example, the relative probability of different kinds of verb complements and adjuncts: direct object, infinitive VP, embedded clause, and prepositional phrase. The diagram also clearly shows the *recursive* nature of English: The eye naturally recognizes, for example, the appearance of NP at different hierarchical levels. More problematically, the diagram represents English as being context-free, in that the probabilities of expansions of a constituent are the same regardless of its syntactic context. In fact, English is *not* context-free in this sense: For example, a subject NP is much more likely to expand to a pronoun than an object NP. In this sense, the diagram oversimplifies the true probabilities of English syntax.

Reference: Marcus, M., Santorini, B., & Marcinkiewicz, M. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19:313—330.*

Table 1. Syntactic rules and their probabilities.

Parent	Expansion	Count	P(Par->Exp)
NP	NP PP	43862	.192
	D N	42193	.185
	N	32758	.143
	ProN	20932	.092
	PropN	15897	.070
	Adj N	15512	.068
	PropN PropN	14881	.065
	D Adj N	13840	.061
	NP N	11299	.049
	NP SBAR	9287	.041
	N N	8104	.035
VP	V NP	30694	.265
	V VP	27964	.242
	V PP	14296	.124
	V	11213	.097
	V SBAR	11213	.097
	V Sinf	10518	.091
	V NP PP	9832	.085
PP	P NP	103708	1.000
S	NP VP	56066	.843
	VP	10459	.157
SBAR	S	9731	.544
	WHNP S	8162	.456
ADVP*	Adv	17331	1.000
Sinf	TO VP	11932	1.000

*The expansion ADVP->Adv could not be included, since no higher-level expansion includes ADVP.

Table 2. Symbols used and their Treebank equivalents and descriptions.

Symbol	TB symbol(s)	Description
A	JJ	adjective
D	DT	determiner
N	NN, NNS	noun
NP	NP	noun phrase
P	IN	preposition
PP	PP	prepositional phrase
ProN	PRP	pronoun
PropN	NNP, NNPS	proper noun
S	S	clause/sentence
SBAR	SBAR	parent of S
Sinf	S/VP	infinitive clause
TH	IN	"that"
TO	TO	"to"
V	VB, VBD, VBP, etc.	verb
VP	VP	verb phrase
WH	WHNP	relative pronoun