MTA SZTAKI

A FINITE AUTOMATON FOR

THE ENGLISH AUXILIARY SYSTEM

András Kornai

II/47

WORKING PAPER

Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intézete
Computer and Automation Institute, Hungarian Academy of Sciences
Institut für Rechentechnik und Automatisierung der Ungarischen Akademie der Wissenschaften
Исследовательский Институт Вычислительной Техники и Автоматизации Венгерской Академии Наук

H 1502 Budapest, XI. Kende utca 13-17. * Hungary * Ungarn * Венгрия *

A FINITE AUTOMATON FOR THE ENGLISH AUXILIARY SYSTEM

One of the primary aims of Gazdar, Pullum and Sag /1980/ in writing "A Phrase Structure Grammar of the English Auxiliary System" was to show that 'the intricate and complexly interacting facts of the English verbal auxiliary system cannot be said to provide any support for the postulation of transformations'/p 100/. Their paper /henceforth EAS[1]/ however, leaves the reader with the impression that a rather formidable apparatus utilizing some fifteen binary features, metarules, rule number features etc. is essential in describing these facts, so that a relatively straightforward transformational treatment is still preferable on grounds of simplicity. In this squib we take a closer look at EAS and show that its substantive content can just as well be expressed by a finite automaton of only eleven states and thereby demonstrate that even the apparent simplicity of, say, the Affix Hopping analysis cannot serve as a justification for a transformational component. By the same token, the mere existence of the EAS analysis does not justify the sophisticated GPSG apparatus and we will show that this is independent of the particular choice of language.

EAS utilizes a two-bar system, and the traditional S → NP VP rule is formulated as $\langle 1, \bar{\bar{V}} \rightarrow \bar{\bar{N}}_{[\alpha]} \bar{V}_{[\alpha]} \rangle$ [2], where $[\alpha]$ is some combination of agreement features[3]. This notation suppresses two all-important conventions: the Archi-Rule Convention or ARC /we coined this name after the terms archiphoneme and archisegment/: 'Where features are left unspecified [...] the node-label stands as a variable ranging over permissible combinatio[ns] of the unspecified features'/p 4/ and the Head Feature Convention or HF[C]

-----

1. In spite of its considerable impact, EAS has never been published /to the best of our knowledge/. Rule and page numbers refer to the September 1980 version which had been widely circulated.

2. We deviate from the original notation in two respects. Limitations of space and the abscence of a fully developed system in EAS make it impossible to discuss the semantics associated with the rules, so we regretfully omit the third member of GPSG triples. We also employ the traditional 'rewriting' notation, instead of the one reflecting the 'node admissibility' interpretation of PSG rules, but this is not a really important modification: for a detailed discussion of this subject see Kornai /forthcoming/

3. In what follows we will disregard this aspect of the analysis, partly because the EAS treatment of agreement is rather vague and partly because it gives rise to a false generalization namely that there is some sort of person-number agreement in English imperatives.

In a rule rewriting X^n, the smallest X^i on the right side has to carry all the feature specifications associated with X^n /if X^i exsists, unique, is not otherwise specified, $i \leq n$ and X is a major category/[4]

This way, we have $S \rightarrow \bar{\bar{N}}\ \bar{V}_{[+FIN]}$, $Q \rightarrow \bar{\bar{N}}\ \bar{V}_{[INV]}$, $I \rightarrow \bar{\bar{N}}\ \bar{V}_{[IMP]}$ and so on where S, Q and I are start symbols for generating tensed sentences, questions and imperatives respectively. Since S $/= \bar{\bar{V}}_{[FIN]}$, see p 19/, Q $/\stackrel{d}{=} \bar{\bar{V}}_{[INV]}/$ and I $/\stackrel{d}{=} \bar{\bar{V}}_{[IMP]}/$ are the only $\bar{\bar{V}}$s that can be rewritten as sentences in EAS we can disregard all other instances of rule 1. Moreover, the specific question and imperative rules in EAS override the Q and I instances of rule 1, and the introducti of S, I and Q wil enable us to dispense with the two-bar level of EAS entirely. The cornerstone of EAS is the following rule system /pp 19-21/:

$\langle n, \bar{V}_{\left[\begin{smallmatrix}+AUX\\ \alpha\end{smallmatrix}\right]} \longrightarrow V_{[n]}\ \ \bar{V}_{[\beta]} \rangle$, $\langle n,\ Q \rightarrow V_{[n]}\ \ \bar{\bar{N}}\ \ \bar{V}_{[\alpha]} \rangle$ where

n	α	β	$V_{[n]}$ membership	
2	+FIN	+BSE	can, may, must, will, could etc.	
3	+FIN	+BSE −AUX	do	
4	+ASP	+PSP	have	
5	+ASP+COP	+PRP	be	
6	+COP	+PAS	be	
7	+INF	+BSE	to	/ only the $\bar{V}$ rule /
8a	+FIN	+INF	ought	
8b	+FIN+COP	+INF	be	
9	+COP	+PRD	be	

and α is −ASP −COP −PRD −PAS −INF and −NUL unless otherwise specified. /The Q rules have the same number because originally these were introduced by a metarule SAI which does not generate Q-rule #7./

Of the 32768 combinations made possible by the fifteen binary features /GER, IMP, INV, NEG and the ones already mentioned/ only 94 are actually permitted by the Feature Cooccurance Restrictions or FCRs /p 17/ In our opinion this points to serious defects in the feature system and

---

4. This definition is equivalent to but not identical with the original one on p 5 of EAS. Although reformulations of this kind play a crucial role in this paper and it would be desirable to demonstrate their equivalence to the original formulation in each and every case, we can at best give hints how equivalence is to be proved : such demonstrations /apart from being extremely lengthy if rigorous enough/ would hardly benefit readers not posessing a copy of EAS anyhow. These readers, however, can reconstruct the EAS rule system from this paper rather faithfully.

in this case a reanalysis is more useful than a straightforward presentation of the original.

Since $\bar{\bar{V}}$[+NUL +NEG] = not /by rule 13 and rule 15 on p 42/ and $\bar{\bar{V}}$[+NUL -NEG] is simply the empty string, NUL can be easily dispensed with./ although this necessitates minor modifications in the format, but not the content of certain rules /. FCR 1 states that of the eight features FIN BSE GER PRP PRD PAS INF and PSP at least and at most one has positive value for any $\bar{\bar{V}}$ /p 14/, therefore we can simulate the effect of these by only three /binary/ features A, B and C./We will take FIN as [-A-B-C] ,BSE as [-A-B+C] ,GER as [-A+B-C] ,... and finally PSP as [+A+B+C]. The ARC and the HFC will act on the new features in the desired manner./ The features IMP and INV can also be eliminated (because +IMP necessitates +BSE /FCR 5/ and +INV necessitates +FIN+AUX /FCR 4/ in favour of BSE and FIN respectively) On the $\bar{V}$ level this can be done without altering the generative capacity of the rules directly, and on the $\bar{\bar{V}}$ level it was precisely because we wanted to dispense with IMP and INV that we had to introduce the symbols I and Q.

This leaves us with a much leaner feature system where less than half of the possibilities are 'accidental gaps' so there is no way to eliminate more features directly.The only cross-categorizing feature in this system is NEG : both values of it can appear freely in any /otherwise permitted/ combination.Thus, /homomorphically/ deleting this feature and modifying the rules accordingly will simplify the grammar by a factor of two.[5] Therefore we will neglect this aspect of EAS but we will have to show that such a step does not really affect the point we intend to make. We emphasize here that the EAS treatment of negation is impeccable / the fact that NEG is a 'true' feature corroborates this / and we took this step solely to make a more convenient presentation possible.This way we have a six-member feature system : in what follows [$\alpha\beta\gamma\delta\varepsilon\zeta$] will stand for [$\alpha$A βB γC δASP εCOP ζAUX]/in this order/ and a zero in some position means that the feature in question is unspecified.In this notation FCR 2 /originally stated as [+PRP] ⇒ [-ASP] / will look like *[-+++00] and FCR 3 is *[++++-0] .We also have three originally unnumbered FCRs /see p 17/ , namely *[000+0-] , *[0000+-] and *[+-+00+] these five taken together will account for the gaps in the system.

5.The actual gain is somewhat less. For a discussion of homomorphisms see Kornai/forthcoming/.

/ If we take it into account that [000--+] combinations can be dispensed with, a more radical reanalysis of the EAS system utilizing only five features /and the original NEG/ is also possible, even one in which the minor descriptive problems of EAS can be remedied without any loss in simplicity./ The morphological feature assignment rules [+0+000] → [en] and [-+0000] → [ing] /p 18/ will be built into the finite automaton by allowing for more than one morpheme on the edges. Moreover, we will allow for more than one label on one edge, but neither of these 'shorthand' devices affects the power of Finite State /FS/ description.

Rule 10 of EAS expands $\bar{V}$[+--000] as $\bar{\bar{N}}$, $\bar{\bar{A}}$ or $\bar{\bar{P}}$ in sentences like 'John is a fool/quite competent/in need of help' /the role of ASP, COP and AUX is entirely unclear here/, and rule 11 introduces the 'do so' construction: ⟨11, $\bar{V}$[-AUX] → V[11] so⟩ where do is the only member of V[11]. In effect, rule 11 is just $\bar{V}$[00000-] ⟶ do so, but we have to keep feature percolation in mind. As the relevant slash category rules are not discussed in the paper and the analyses developed elsewhere /e.g. Gazdar 1981/ are incompatible with EAS we have to omit topicalization /rule 12/. Rule 14 is $\bar{V}$[----00+] → {may/could} not $\bar{V}$[--+000] and the HFC applies here as well.[6] Since we have only [-NEG] $\bar{V}$s the three original imperative rules /p 36/ can be collapsed into one: I → {do / dont (N̄)} $\bar{V}$[--+000] and here the HFC does not apply. The 'VP-Deletion' /VPD/ metarule /p 42/ makes it possible to drop the $\bar{V}$ on the right side of any rule expanding $\bar{V}$[+0000+] or $\bar{V}$[0-000+] as V$\bar{V}$ /or to substitute it by not/ --this will be indicated here by the adverb too. Although EAS provides 'a metarule to handle the facts about sentential adverb placement in a variety of English described by Jackendoff/1972/ which only permits the adverb after the first auxiliary verb', such a variety does not exist /cf Jackendoff 1977 ch. 4.4 esp. fn. 15/ and as the real facts of adverb placement cannot be handled with the EAS grammar any concise way, this rule will be ignored here. In spite of the fact that such a rule is only implicit in EAS we assume that main verbs are to be generated under $\bar{V}$[000---] /with the possible exception of $\bar{V}$[+PRD]/.

The reader can verify for himself that the finite automaton on Fig.1 is a faithful rendering of the EAS grammar : this task is facilitated by the relevant $\bar{V}$ feature combinations and rule numbers displayed on the states and edges respectively.

------

6. The main rule involving negation that we cannot include in the discussion is ⟨13, $\bar{V}$[-FIN+NEG] ⟶ not $\bar{V}$[-NEG]⟩ .

A close inspection of this automaton will reveal several minor descriptive flaws in EAS : for instance it generates

*John is do so /rules 1,8 and 10/

*/Bill is to leave and/ John is to too /rules 1,8 and VPD/

/The first of these can be easily excluded by stipulating *[++-OO-] ,but the second points to a serious problem in regarding <u>to</u> an auxiliary /cf p 22/./ Moreover,this automaton enables us to take stock of the whole grammar: for instance a dubious sentence type like

Don't have been eating! /rules 16,4 and 5/

which might be argued for in certain contexts /cf. p 97/ now appears as only one instance of a mistaken generalization.Since the above sentence, 'John is to have been eating' and 'John can have been eating' are all derived through the state BSE 2, the reader can see at a glance that the crux of the matter is the lack of temporal distinctions in the phrases generated under V[BSE] and can solve the problem accordingly.Another example is that of passives: it is obvious from the automaton that EAS treats sentences like 'John is a fool' and 'John is eaten by a bear' on a par: there is simply no derivation that can yield one of these but not the other /by taking the alternate last step/[7]

This sort of transparency was highly valued traditionally /cf. e.g. Bloomfield 1945/ but the possibility of computerized testing clearly diminishes its importance today./Unfortunately most of the work now being done in Government-Binding is far too fuzzy to admit for computerized testing or, indeed for any attempt at falsification./

FS description, however, is also transparent in the acquisition sense: given an arbitrary FS language and a finite automaton there is an algorithm to decide whether the automaton accepts the language or not.This is not true in the CF case: given a CF language /for instance the set of all conceivable strings/ there is no algorithm in general to decide whether an arbitrary CF grammar generates this language or not. This way, GPSG

7.As a matter of fact,the "like coordination" and "consider" tests proposed in EAS /p 26/ to justify the grouping of N, A and P under V +PRD cannot distinguish between V +PAS and the above categories.Witness:

Kim is tormented by nightmares and frightened of them /V +PAS and P/

Kim is a good striper and paid well by the <u>Sands</u> /N and V +PAS/

Kim is dumb and ignored by all her teachers /A and V +PAS/

and

Kim considered Sandy spoiled by her grandparents/a nice person/etc.

cannot claim explanatory adequacy as long as the class of languages that can be described in this framework equals the CF class, and this probably applies to the Government-Binding theory even more forcefully--unfortunately GB is not presented rigorously enough to make such an evaluation possible./For the proofs of the mathematical theorems evoked in this paper see e.g. Salomaa 1973/

Pullum and Gazdar /1982/ provides 'a simple proof that an infinite language with long-distance dependencies and syntactic concord over unbounded domains can be an FS language' /p 474/ and the reader can construct the finite automaton incorporating the EAS treatment of negation himself : all one has to do is to take another copy of this automaton /to model the [-NEG] feature combinations/ and to add and delete certain lines.As for topicalisation and adverb placement it suffices to note that basically all the EAS rules are of the form $N \rightarrow \alpha M$ where N and M are nonterminals /$\overline{\overline{V}}$s and $\overline{\overline{V}}$s/ and α is a string of terminals /all other categories employed in EAS are left unanalyzed and therefore can be regarded terminals/ and a generative grammar with such a restricted rule format is always equivalent to a finite automaton.

One is tempted to say that the whole demonstration was based on this observation or even on the fact that the English auxiliary system is an essentially finite corpus and therefore lends itself to FS description no matter how rules of some grammar describing it might look.This is not the case, however: the phenomenon illustrated by this paper namely that FS description is a 'free rider' on CF grammars is independent of corpus and indeed independent of language as long as the language in question is a natural one i.e. it is subject to the memory limitations of its human users.In that case, the Depth Hypothesis of Yngve /1961/ applies therefore an adequate CF grammar cannot be self-embedding and the stringset generated by it can be accepted by a finite automaton.[8]

8.For a more detailed argument and independent motivation for the Depth Hypothesis see Kornai /forthcoming/ where it is also demonstrated that natural languages /as stringsets/ must be finite-state no matter what position one takes on their CF-ness and self-embedding.

References

Chomsky,N.: Syntactic Structures.1957 Mouton,The Hague

Gazdar,G.J.M.,G.K.Pullum and I.Sag: A Phrase Structure Grammar for the English Auxiliary System. Unpublished manuscript. 1980

Gazdar,G.J.M.: Unbounded dependencies and coordinate structure. Linguistic Inquiry 12 /1981/ 155-184

Jackendoff,R.: $\bar{X}$-syntax: a study of phrase-structure,1977 MIT Press, Cambridge,Mass.

Kornai,A.:"... some version of the $\bar{X}$-theory".To appear.

Pullum,G.K. and G.J.M.Gazdar: Natural languages and context-free languages.Linguistics and Philosophy 4 /1982/ 471-504

Salomaa,A.: Formal Languages.1973 Academic Press,New York

Yngve,V.H.: The Depth Hypothesis. Proceedings of Symposia in Applied Mathematics XII /1961/ 130-138

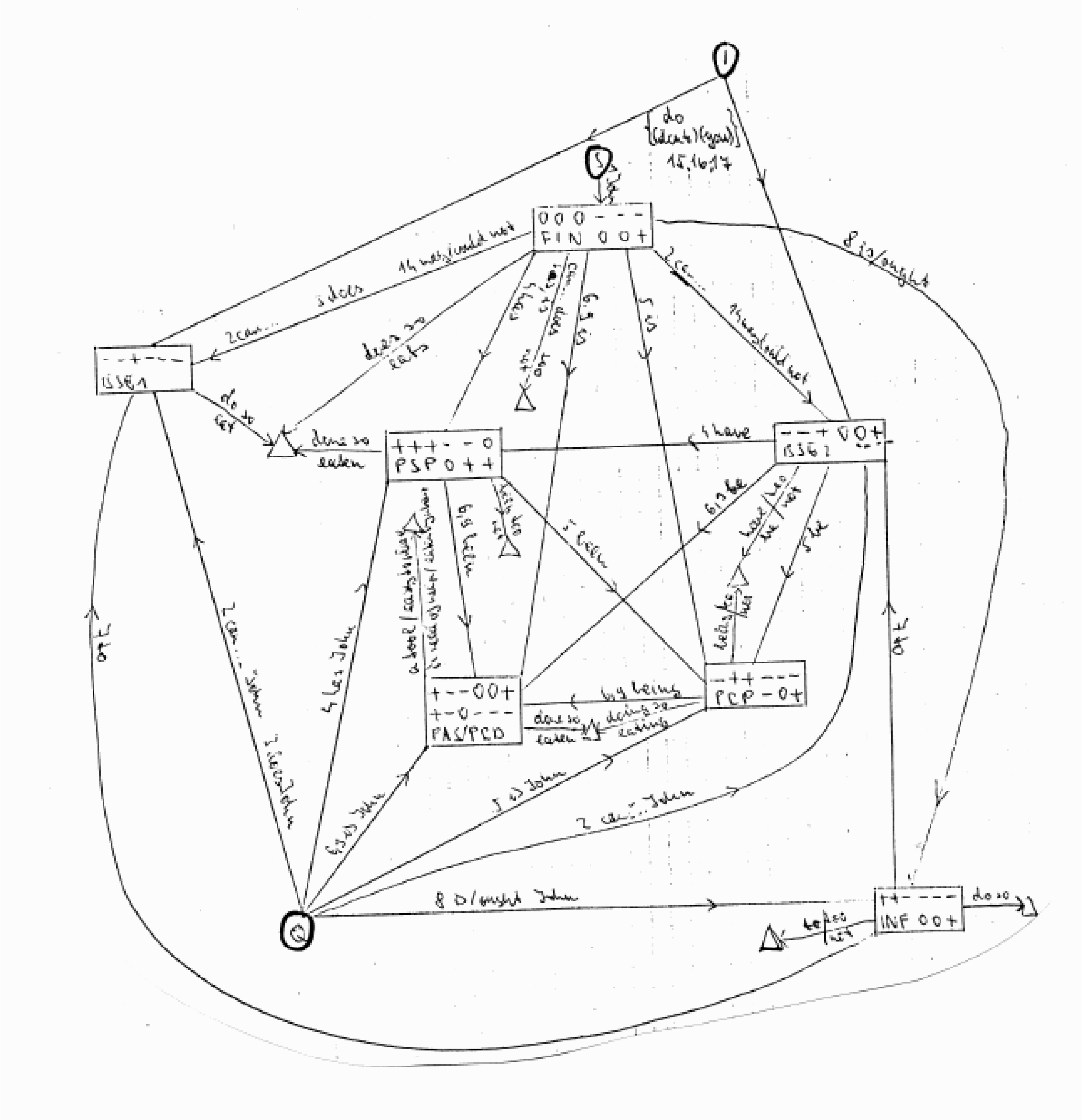

Fig. 1.

Initial states are marked by a single inscribed letter, and the triangle denotes the /one and only/ accepting state.