Richard Sproat: A Computational Theory of Writing Systems

Cambridge University Press, Cambridge, UK.
[hardcover, $59.95]

Markus Walther*
(Panasonic Speech Technology Laboratory)

1. SYNOPSIS

This is a remarkable book. Written by an author who has previously shown with his 1992 classic, Morphology and Computation by MIT Press, that he is well-versed in both linguistic detail and computational matters, he strikes that rare balance once again with his newest work. The book is to be understood against the author's extensive background in developing text analysis components for text-to-speech synthesis (TTS) at AT&T Bell labs. What drives this book is the conviction of the author that—starting from an orthographically represented text—the problem of modelling the task of reading aloud within a TTS system has a great deal to say for a model of the same task, as performed by the human reader.

After a brief introduction to text-to-speech conversion to the

[*] Markus Walther received his Ph.D. in computational linguistics in 1997 from the university of Duesseldorf, Germany. His thesis work concentrated on a constraint-based computational model of prosodic morphology, developing a formally sound a-templatic theory of non-concatenative morphology. Postdoc research at the University of Marburg, Germany, thereafter has broadened that interest to other phenomena, in particular reduplication and truncation (see www.markus-walther.de). Recently, Markus Walther has started to work on text-to-speech systems at the Panasonic Speech Tech Lab, Santa Barbara, USA.

uninitiated, chapter 1 ("Reading Devices") rather quickly proceeds to some formal preliminaries and an axiomatic presentation of the core of Sproat's theory, before the two central claims of his model of reading aloud are formulated (p.16): **Regularity** – "The mapping from the Orthographic Relevant Level (ORL) to the spelling level itself is a regular relation" (in the mathematical sense of formal language theory), and **Consistency** – "The ORL for a given writing system (as used in a particular language) represents a consistent level of linguistic representation." On p. 10 this central concept of ORL is defined as the linguistic level where correspondences between linguistic elements and their orthographic expression are most succinctly stated. Sproat's exemplification here is Russian, whose ORL he claims to be the morphological level, i.e. a level where morphologically related forms are typically represented consistently, abstracting away from phonological changes. A word like <goroda> can be pronounced with initial stress to mean 'of a city' or with final stress to denote 'cities' – but stress and its accompanying vowel reductions are not marked, hence orthographical coverage of the phonological level is incomplete. This deeper-than-surface-phonemic level contrasts with neighbouring Belarussian, which does mark these reductions, thus exhibiting a more 'shallow' ORL.

The aforementioned axioms that Sproat puts forward as a formal background essentially take up the temporal-intervals framework that was pioneered by Bird and Klein (1990) for phonology. They specify that – unless lexically overridden – immediately preceding elements on the linguistic level will appear concatenated at the spelling level (English /bo/ → <bo>, but /ks/ → <x> lexically), that in domination relationships at the ORL it is the dominated element which will be spelled out by default, and that for temporally overlapping linguistic elements the spellout is the concatenation of the individual ORL-to-spelling mappings. Sproat decomposes this overall mapping into a set of graphic encoding rules and a set of autonomous spelling rules, both of which must be regular under his first claim (p.18). His second claim is made more precise by assuming a cascade of N levels of representation mediated by classic rewrite rules—**Consistency** then amounts to picking the ORL from a level i, $1 \leq i \leq N$. (While Sproat leaves open the possibility of a nonderivational, constraint-based recasting of these notions, he asserts that some languages are most naturally modelled by referring to levels of representation). The chapter concludes with terminology and conventions, and presents a self-contained introduction to regular relations and their corresponding computational devices, finite-state transducers (which continue to rise in importance in contemporary computational linguistics).

Chapter 2 ("Regularity") defends the first claim in more detail. First, a generalization from one-dimensional left-to-right concatenation to the two dimensions of the plane is carried out, allowing formal description of e.g. the subcomponents of Chinese characters with the help of planar finite-state automata that can catenate components in leftward, rightward, downward, upward and surrounding fashion. In that context, a third claim of **Locality** is made, namely that deviation from the text-inherent global catenation direction (e.g. left-to-right) only occurs within a graphic unit corresponding to a 'Small Linguistic Unit' (e.g. the syllable in Chinese). Further cases from Korean Hangul, Devanagari and Pahawh Hmong scripts are treated, before an extensive set of placement rules for Chinese character components is presented as a case study. Sproat is honest enough not to leave out a counterexample from Ancient Egyptian, where plural marking by orthographic word doubling is *not* mirrored in linguistic reduplication (an ordinary suffix was used) —this in principle problematic because arbitrary string copying is wellknown to be a non-regular relation. However, the ensuing conclusion that his theory predicts this as marked, hence rare, is incorrect insofar as no gradient or probabilistic element is found in the theory, as required for rarity or relative markedness.

Chapter 3 ("ORL Depth and Consistency") adresses the second claim by taking up Russian and Belarusian in more detail, before presenting a most interesting comparison of deep versus shallow ORL for English.

Sproat challenges the standard position held since Chomsky & Halle (1968) of a 'deep' ORL by giving URs of 1,169 words chosen to map the territory where SPE excelled—and showing that a 'shallow', surfacy alternative does remarkably well under its own set of rewrite rules, which are listed in an appendix. Again, a counterexample is discussed: Serbo-Croatian appears inconsistent to **Consistency** insofar as voicing assimilations show up in the orthography, but /d/ before (ordinary & palatalized) /s/ retains spelling <d>. To investigate the underlying phonetic reality, Sproat actually conducts a pilot experiment with a native speaker, which surprisingly shows that in fact the orthography rather faithfully reflects the observation that [-voice] assimilation is gradient and *least complete* precisely before the two fricatives, both confirming **Consistency** and contradicting the traditional analysis.

Chapter 4 ("Linguistic Elements") investigates "the range of linguistic elements than can be represented by written symbols in the world's writing systems" (p.131). The author criticizes Gelb's, Sampson's and DeFrancis' tree-based classification, devising his own two-dimensional chart with dimensions 'Amount of Logography' (e.g. English is less logographic than Japanes) and 'Type of Phonography' (e.g. W.Semitic is Consonantal, English is Alphabetic, Chinese is Syllabic). A case study of Chinese follows, where Sproat shows that his axioms correctly predict a duplication of semantic radicals across both components of a disyllabic morpheme, apparently the first explanation of its kind. Of the further examples, I found reduplication markers to be of particular interest. Khmer, Malay and Bahasa Indonesia are among those who mark repetition of preceding material, the latter two using a (sometimes raised) "2". Contrary to Sproat ("earlier forms", "was used"), however, this definitely is a contemporary marking device: e.g. searching for <orang2> `person (pl.)' on Indonesian (.id) web pages alone produces 242 hits, even showing recent forms like <foto2> as a byproduct. This is of interest as it could present another potential counterexample to **Regularity**, due to reduplicative copying. Sproat solves this by assuming reduplicative morphemes to be already marked at the ORL level, rendering mappings such as orang $[_{\text{copy}}\text{orang}]_{\text{copy}}$ → orang2 an easy task. However, he does not discuss that the inverse mapping is potentially problematic under productivity: if a new word like <fax2> is read aloud for the first time, native speakers will pronounce this *only* as /faksfaks/, even though their mental lexicon does not contain the word yet, countrary to Sproat's apparent reliance on a complete lexicon here. [Sproat (p.c.) explains that he would delegate reduplicative copying to a more powerful morphological module M. In practice, this would reduce the ORL mapping to orang2 → orang$[_{\text{copy}}$ "any string" $]_{\text{copy}}$ for our case, something which takes only regular power to accomplish, and "any string" would then be restricted to "orang" by M, which is separate from ORL. This could be a workable solution, once some technical details are solved.]

Chapter 5 ("Psycholinguistic Evidence") proceeds to "see if there is any support in the psycholinguistic literature for some properties of the model" (p.163). Here, Sproat focuses on the predicted crosslinguistic sameness of the nature of the relation between orthography and linguistic form, and the prediction of a dual route, i.e. "an additional rule-based path to pronunciation that bypasses the lexicon" (p.164). His review of the literature finds these two to be well-supported, not withstanding some connectionist counter-proposals that deny dual-route. He dissects one of the latter - the Seidenberg-McClelland model of 1989—and classifies it as a "toy system" that is not tested on realistic data sets.

A final chapter 6 titled "Further Issues" looks into adaptation of writing systems to a new language (Manx Gaelic is the example here), spelling reforms (the case of the 1995 reform for Dutch), the relation between written numerals and their fully spelled-out number names, the orthographic encoding of abbreviations and the possibility that written language is on a par with spoken language. Of these the discussion on numerals was the most fascinating topic to the present reviewer, as Sproat sets forth to devise a general strategy for first factorizing number strings like 3684 into a factors-and-powers-of-ten representation, and then mapping to the number words that exist in a given language. Unlike, say, English, the case of present-day Malagasy is a hard one for Sproat, though, because this language reverses the logical order, amounting to "four and eight+ten and six+hundred and three thousand". Since only a finite set of strings can be reversed (by sheer enumeration, i.e. at great cost) under **Regularity**, he must assume a low-level processing step of reading-direction reversal to handle arbitrarily long numbers in this language (he does not cite experimental evidence to support this prediction), though he has a point in saying that pre-19th century Arabic-script writing in Malagasy would not have had that peculiarity.

2. CRITIQUE

While the overall quality of this scholarly work is impressive, good pointers to the literature are given throughout, and relevant discussion is nowhere suppressed, a few points are nevertheless worthy of criticism. A general one is Sproat's employment of "soft" notions like Orthographically Relevant Level, **Consistency**, Small Linguistic Unit, which are not rigidly defined. Consider **Consistency**, where one has to pick a level in a N-rule cascade as the ORL. Since URs themselves are not strictly empirically derived, one can always explain away exceptions to **Consistency** by lexical marking (as done e.g. on p.78f,89), making the concept hard to falsify. Tightening up this concept and making it fully empirical would probably require something like minimum-description-based automatic learning of both rules *and* exception features, and/or replacement of URs by phonetic SRs. Also, it is not clear to this reviewer why a special type of planar automata had to be devised to cope with two-dimensional character arrangements (p.40f)—could not the same task be accomplished by a string-linearized *description* of those arrangements, thus reverting to standard finite-state automata? Note that planar automata as defined by Sproat need an additional interpretation step anyway, since their output is also to be understood as a description; it does not give concrete pixels on a page. As in every book, there are also a few cases where the logic seems quirky, e.g. when Sproat criticizes Wang's analysis of within-character placement of components in Chinese. He dismisses Wang's featural decomposition of placement as overly powerful because it allows a component to be placed *inside* another one, whereas according to Sproat such placement only occurs in fossilized forms. But traditional logic in the field would rather cite this as support that Wang is on the right track, since there apparently *was* a stage in the language where this could happen. Since Sproat's theory is essentially possibilistic rather than probabilistic, ruling out cases on appeal to markedness/infrequency/ unproductivity seems a bit odd. Also, there is the fact that several instances of potential crosslinguistic violations of **Regularity** (recall in particular the reduplication cases; one could also cite balanced French/German-type quotes-in-quotes-quotes, which are akin to the non-regular formal language $a^n b^n$) had to be explained away by appeal to different devices. This makes the present reviewer wonder whether not in fact the "true" state-of-affairs might be better described by some mildly non-regular formal language framework that is somehow probabilistically weighted to reflect the prevailing regular cases, notwithstanding the undisputed utility of the finite-state paradigm in natural language engineering. Speaking of the few outright errors and omissions, appendix A of ch.3 comparing deep and shallow ORL for English appears noteworthy. A feature [+db] is introduced with no rules referring to it. As a consequence, e.g. the claimed mapping of the ORL for 'allophone' to its orthographic representation is not derivable (p.99), with the only rule for translating /l/ being # 30: l → <l> (p.129). This is unfortunate in the light of the claim (p.85) that rules have been tested against ORL-spelling pairs (fn.20 specifically advertises AT&T's FSM software as the one used for concrete testing). [Sproat (p.c.) confirms successful computerized testing, so the incompleteness appears to be only in the book version.] Similarly, no rule refers to secondary stress, yet it is included in both ORLs. Also, no rule for '+' (morph.boundary) could be found despite *rhythmic+s,spheric+s,vertic+es*, one apparently would need to add iz → <es> / + __ #. Then there are undiscussed rule types like dZ → <g> / __

(<i> | <e> | <y>) (p.130), which—by referencing the output level in their right context—could be interpreted as two-level rules in the sense of Koskenniemi (1983). Independently, this reference to output would be the only way words like ORL 'dZin → ORTH <gene> (p.109) could be derived, where the vowel /i/ has no lexical marking <i> in Sproat's listing. This important detail clearly should have been mentioned (two-level rules are mentioned only for non-crucial orderings, p.69, fn.2., and no notational conventions are given for them). But, overall, the book has much to recommend it for, and should be seen as an important contribution to the modern study of crosslinguistic writing systems for some years to come.

BIBLIOGRAPHY

Bird, S. & Klein, E. (1990). Phonological events. *Journal of Linguistics 26*, 33-56.

Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*. Harper & Row.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. University of Helsinki.

Seidenberg, M. & McClelland, J. (1989). A distributed, developmental model of visual word recognition and naming. *Psychological Review, 96,* 523-568.

Sproat, R. (1992). *Morphology and Computation*. MIT Press.

Markus Walther
Panasonic Speech Technology Laboratory
3888 State Street, Suite #202
Santa Barbara, CA 93105, USA
Email: mwalther@stl.research.panasonic.com
WWW: http://www.markus-walther.de