

计算语言学
与语言科技
原文丛书

CAMBRIDGE

A COMPUTATIONAL THEORY
OF WRITING SYSTEMS

文字书写系统的
计算理论

作者 Richard Sproat
导读 陆 勤



北京大学出版社
PEKING UNIVERSITY PRESS

著作权合同登记 图字 01-2009-4468

图书在版编目(CIP)数据

文字书写系统的计算理论: A Computational Theory of Writing Systems/史伯乐(Sproat, R.) 著. —北京:北京大学出版社, 2010. 8

(计算语言学与语言科技原文丛书)

ISBN 978-7-301-17155-4

I. 文… II. 史… III. 自然语言处理—英文 IV. TP391

中国版本图书馆 CIP 数据核字(2010)第 075891 号

A Computational Theory of Writing Systems, first edition (ISBN: 978-0-521-03422-1) by Sproat first published by Cambridge University Press 2000.

All rights reserved.

This reprint edition for the People's Republic of China is published by arrangement with the Press Syndicate of the University of Cambridge, Cambridge, United Kingdom.

© Cambridge University Press & Peking University Press 2010

This book is in copyright. No reproduction of any part may take place without the written permission of Cambridge University Press and Peking University Press.

This edition is for sale in the People's Republic of China (excluding Hong Kong SAR, Macau SAR and Taiwan Province) only.

此版本仅限在中华人民共和国境内(不包括香港、澳门特别行政区及台湾地区)销售。

书 名: 文字书写系统的计算理论

著作责任者: Richard Sproat 著

责任编辑: 白雪 李 凌

标准书号: ISBN 978-7-301-17155-4/H · 2496

出版发行: 北京大学出版社

地 址: 北京市海淀区成府路 205 号 100871

网 址: <http://www.pup.cn>

电子邮箱: zpup@pup.pku.edu.cn

电 话: 邮购部 62752015 发行部 62750672 编辑部 62753334

出版部 62754962

印 刷 者: 世界知识印刷厂

经 销 者: 新华书店

787 毫米×1092 毫米 16 开本 17.25 印张 287 千字

2010 年 8 月第 1 版 2010 年 8 月第 1 次印刷

定 价: 34.00 元

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究 举报电话: 010-62752024

电子邮箱: fd@pup.pku.edu.cn

丛书前言

计算语言学(Computational Linguistics, CL)在语言科学与信息科学的研究领域扮演关键性的角色。语言学理论寻求对语言现象规律性的揭示与完整的解释。计算语言学正好提供了验证与应用这些规律与解释的大好机会。作为语言学、信息科学乃至心理学与认知科学结合的交叉学科,计算语言学更提供了语言学基础研究与应用研究的绝佳界面。事实上,计算语言学与人类语言科技(Human Language Technology, HLT)可以视为一体两面,不可分割。

计算语言学研究滥觞于上世纪五六十年代的机器翻译研究。中文的相关研究也几乎同步开始,1960年起在柏克莱加州大学研究室,王士元、邹嘉彦、C. Y. Dougherty 等人已开始研究中英、中俄机器翻译。他们的中文计算语言学研究,可说是与世界最尖端科技同步的。中国国内中俄翻译研究也不遑多让,大约在上世纪 50 年代中期便已开始。可惜的是,这些中文相关早期机器翻译研究,由于硬件与软件的限制,没能延续下来。中文计算语言学研究比较有系统的进展,还要等到 1986 年;海峡两岸在同一年成立了两个致力于中文计算语言学基础架构建立的研究群。北京大学的计算语言学研究所在朱德熙先生倡导下成立,随后一段时间由陆俭明、俞士汶主持。而台湾“中研院”的中文词知识库小组,由谢清俊创立,陈克健主持,黄居仁 1987 年返台后加入。

中文计算语言学的研究,20 余年来已累积了相当可观的成绩,重要研究领域与议题中都有可观的研究成果,华人计算语言学者也渐渐在国际学术界崭露头角。随着世界经济转向知识密集产业,跨语言跨文化沟通与知识整合成为知识产业的关键,语言科技的发展日渐成为国际主流。在这个有利发展的大环境下,我们相信,中文计算语言学与华人计算语言学学者的成绩,将会百尺竿头更进一步,进入计算语言学学术核心,并产生把握学科动态、引领学术走向的大师。

回顾计算语言学研究在过去二十年的蓬勃发展,统计模式的引入应该是最主要的原因之一。但二十年后学界也开始看到了统计模式的局限,因此最近几届 ACL 终身成就奖得主,不约而同地大力提倡结合语言学理论与概率模型的研究,来提升计算语言学研究的层次,以寻求新的突破。

回顾中国国内的计算语言学发展,来自计算机科学的贡献多于语言学的贡献。这在理论与概率模型整合研究的大趋势下,不免令人忧心。这也许可以部分归咎于英文研究专著获得不易。国内较易取得期刊或会议论文,但由于篇幅的限制,往往无法对理论做深入完整的阐述,因此也导致国内年轻学者,长于运算而拙于理据。因此,藉由英文专书来弥补不足,巩固研究理据,进而开拓研究视野,是非常重要的一步。

剑桥大学计算语言学原版书系列的引进,就是在上述背景下产生的。本人忝为 Cambridge University Press 所出版的 Studies in Natural Language Processing 系列编辑委员之一,并将于 2010 接任主编。能够将此系列中较重要的几部著作引进国内,责无旁贷。引进原版,不是难事;要真正搭建知识的桥梁,使国内学者与学生开拓研究视野,将原文著作的理论精髓,更多应用于中文研究,则需另加努力。因此,本丛书的特色,是在保留原版的基础上,每本书都邀请一位专家撰写中文导读,其着力点有三:

其一,全书内容简介。导读作者长年浸淫于该领域,对原著能提纲挈领,切中肯綮,并提供相关研究背景。可助读者更准确地掌握并吸收该书的内容。其二,中文相关研究。原作不一定会提到相关的中文研究。由导读专家补充介绍,能搭起理论与中文相关应用的桥梁,从而能够使读者掌握在这个议题进入中文研究的最佳切入点,让相关中文研究的开拓者获得理论的参照和指导。其三,补充原书出版后该领域研究的新发展。现代科技发展迅速,任何经典著作出版后,几乎马上有新的相关研究。因此,在理论架构的脉络中,加上新近发展,使读者能更贴切地掌握研究脉动。全书摘要通常采用文字叙述。而中文相关研究及最新研究发展则分别以文字叙述及延伸阅读书目的方式呈现。延伸阅读书目,使读者可以很快上手,进入相关研究领域,也是本丛书策划者的苦心所在。可以说导读是本丛书的亮点,不特为原书增色,亦且增加了不少附加价值。

本丛书的出版,是多方协作的结果。在规划出版的漫长过程中,北大计算语言学研究所俞士汶老师及常宝宝老师提供了无私无悔的支持。香港理工大学,特别是北大一理大汉语语言学研究 centers 与陈瑞端、石定栩、沈阳几位在关键时刻的挹注,也起到了关键作用。当然,整个系列能够顺利出版,离不开有学术眼光和胸襟的北大出版社的支持,而剑桥出版社主管编辑 Helen Barton 从中斡旋,使合约能顺利签订,是必不可少的一环。最后,我要感谢本丛书的国内编委,特别是此次担任导读的各位主笔的辛勤付出,他们为读者搭建了进入学术殿堂的台阶。本丛书的出版,适逢 2010 COLING 国际计算语言学会议在北京举办之际,正象征着国内计算语言学研究与国际的接轨;国内学者风云际会,大展身手,跻身计算语言学的国际舞台,将指日可待。

丛书主编

黄居仁

谨志于香港红磡

二零一零年元月

目 录

导读	F9
Preface	F29
List of Figures	F31
List of Tables	F33
1 Reading Devices	1
1.1 Text-to-Speech Conversion; A Brief Introduction	2
1.2 The Task of Pronouncing Aloud; A Model	6
1.2.1 A Simple Example from Russian	6
1.2.2 Formal Definitions	11
1.2.2.1 AVMs and Annotation Graphs	11
1.2.2.2 Definitions	13
1.2.2.3 Axioms	14
1.2.3 Central Claims of the Theory	15
1.2.3.1 Regularity	16
1.2.3.2 Consistency	19
1.2.4 Further Issues	20
1.2.4.1 Why a Constrained Theory of Writing Systems?	21
1.2.4.2 Orthography and the “Segmental” Assumption	23
1.3 Terminology and Conventions	25
1. A Appendix; An Overview of Finite-State Automata and Transducers	29
1. A. 1 Regular Languages and Finite-State Automata	29

1. A. 2	Regular Relations and Finite-State Transducers	30
2	Regularity	34
2. 1	Planar Regular Languages and Planar Regular Relations	35
2. 2	The Locality Hypothesis	41
2. 3	Planar Arrangements: Examples	42
2. 3. 1	Korean Hankul	43
2. 3. 2	Devanagari	45
2. 3. 3	Pahawh Hmong	47
2. 3. 4	Chinese	48
2. 3. 5	A Counterexample from Ancient Egyptian	54
2. 4	Cross-Writing-System Variation in the SLU	55
2. 5	Macroscopic Catenation: Text Direction	59
2. A	Sample Chinese Characters and Their Analyses	62
3	ORL Depth and Consistency	67
3. 1	Russian and Belarusian Orthography: A Case Study	67
3. 1. 1	Vowel Reduction	68
3. 1. 2	Regressive Palatalization	73
3. 1. 3	Lexical Marking in Russian and Other Issues	76
3. 1. 4	Summary of Russian and Belarusian	79
3. 2	English	79
3. 3	The Orthographic Representation of Serbo-Croatian Consonant Devoicing	89
3. 3. 1	Methods and Materials	91
3. 3. 2	Results	92
3. 4	Cyclicity in Orthography	95
3. 5	Surface Orthographic Constraints	96
3. A	English Deep and Shallow ORLs	99
3. A. 1	Lexical Representations	99

3. A. 2	Rules for the Deep ORL	127
3. A. 3	Rules for the Shallow ORL	129
4	Linguistic Elements	131
4. 1	Taxonomies of Writing Systems: A Brief Overview	132
4. 1. 1	Gelb	132
4. 1. 2	Sampson	133
4. 1. 3	DeFrancis	134
4. 1. 3. 1	No Full Writing System Is Semasiographic	134
4. 1. 3. 2	All Full Writing Is Phonographic	135
4. 1. 3. 3	Hankul Is Not Featural	135
4. 1. 4	A New Proposal	139
4. 1. 5	Summary	143
4. 2	Chinese Writing	144
4. 3	Japanese Writing	154
4. 4	Some Further Examples	160
4. 4. 1	Syriac <i>Syame</i>	160
4. 4. 2	Reduplication Markers	161
4. 4. 3	Cancellation Signs	162
5	Psycholinguistic Evidence	163
5. 1	Multiple Routes and the Orthographic Depth Hypothesis	165
5. 1. 1	Evidence for the Orthographic Depth Hypothesis	168
5. 1. 2	Evidence against the Orthographic Depth Hypothesis	170
5. 2	“Shallow” Processing in “Deep” Orthographies	171
5. 2. 1	Phonological Access in Chinese	171
5. 2. 2	Phonological Access in Japanese	173

5.2.3	Evidence for the Function of Phonetic Components in Chinese	175
5.2.4	Summary	176
5.3	Connectionist Models; The Seidenberg-McClelland Model	178
5.3.1	Outline of the Model	179
5.3.2	What Is Wrong with the Model?	181
5.4	Summary	184
6	Further Issues	185
6.1	Adaptation of Writing Systems; The Case of Manx Gaelic	185
6.2	Orthographic Reforms; The Case of Dutch	192
6.2.1	The 1954 Spelling Rules	193
6.2.2	The 1995 Spelling Rules	195
6.3	Other Forms of Notation; Numerical Notation and Its Relation to Number Names	197
6.4	Abbreviatory Devices	204
6.5	Non-Bloomfieldian Views on Writing	209
6.6	Postscript	212
	Bibliography	217
	Index	227

List of Figures

- 1.1 A partial linguistic representation for the sentence in (1.1). Shown are a phonetic transcription, a prosodic analysis into two intonational phrases (ι) and one utterance (U), accent assignment (\ast), a set of part-of-speech tags, and a simple phrase-structure analysis. Phonetic symbols are IPA. Note that 'MP' means 'measure phrase'. 3
- 1.2 A simple FST implementing the rewrite rules in (1.5). In this example the machine has a single state (0), which is both an initial and a final state. The labels on the individual arcs consist of an *input label* (to the left of the colon) and an *output label* (to the right). Here, capital Roman letters are used to represent the equivalent Cyrillic letters. 17
- 1.3 An acceptor for ab^* . The heavy-circled state (0) is (conventionally) the initial state, and the double-circled state is the final state. 30
- 1.4 An FST that accepts $a:c (b:d)^*$. 32
- 1.5 Three transducers, where $T_3 = T_1 \circ T_2$. 32
- 2.1 $\gamma(\alpha) \rightarrow \gamma(\beta) \downarrow \gamma(\zeta) \rightarrow \gamma(\delta)$. 36
- 2.2 Another figure described by $\gamma(\alpha) \rightarrow \gamma(\beta) \downarrow \gamma(\zeta) \rightarrow \gamma(\delta)$. 36
- 2.3 The five planar concatenation operations: (a) $\gamma(\alpha) \rightarrow \gamma(\beta)$, (b) $\gamma(\alpha) \leftarrow \gamma(\beta)$; (c) $\gamma(\alpha) \downarrow \gamma(\beta)$; (d) $\gamma(\alpha) \uparrow \gamma(\beta)$; (e) $\gamma(\alpha) \odot \gamma(\beta)$. 38
- 2.4 A 2FSA that recognizes Figure 2.2. The labels "R" and "D" on the arcs denote reading direction; "left" on state 0 (the initial state) denotes the position at which scanning begins. 41
- 2.5 A 2FST that maps the expression $\alpha\beta\zeta\delta$ to $\gamma(\alpha) \rightarrow [\gamma(\beta) \downarrow \{\gamma(\zeta) \rightarrow \gamma(\delta)\}]$. 42
- 2.6 The syllables <mos> /mo/ 'cannot' and <cal> /čal/ 'well' in Hankul. 44
- 2.7 A two-word Mayan glyph representing the adjective sequence *yax ch'ul* 'green sacred', after Macri (1996,

	page 179), with the arrangement $YAX \rightarrow [ch'u^{\dagger}li]$. Per convention, the capitalized gloss represents a logographic element, and the lower case glosses phonographic elements.	57
2.8	Three methods of wrapping the virtual tape: (a) standard non-boustrophedon; (b) boustrophedon; (c) inverted boustrophedon.	61
3.1	Waveform and voicing profile for one utterance of <i>gradski</i> 'urban'. The closure for the /d/ is labeled as "dcl", the voicing offset is labeled as "v", and the start of the /s/ is labeled as "s". The voicing profile is the third plot from the top in the second window, labeled as <i>prob_voice</i> .	92
3.2	Barplot showing the proportions of voicing for all samples of underlying /d/ (black bars) versus /t/ (shaded bars).	93
4.1	The taxonomy of Gelb (1963), along with examples of writing systems that belong to each case.	133
4.2	The taxonomy of Sampson (1985).	133
4.3	Featural representation of Korean Hankul, from Sampson (1985, page 124, Figure 19). (Presented with permission of Geoffrey Sampson.)	137
4.4	DeFrancis's classification of writing systems.	138
4.5	A nonarboreal classification of writing systems. On Perso-Aramaic, see Section 6.1.	142
5.1	A model of reading a word aloud, simplified from Besner and Smith (1992).	166
5.2	The Seidenberg and McClelland model of lexical processing (Seidenberg and McClelland, 1989, page 526, Figure 1). Used with permission of the American Psychological Association, Inc.	179
5.3	The implemented portion of Seidenberg and McClelland's model of lexical processing (Seidenberg and McClelland, 1989, page 527, Figure 2). Used with permission of American Psychological Association, Inc.	180
5.4	Replication of the Seidenberg, McRae, and Jared (1988) experiment, from Seidenberg and McClelland (1989, page 545, Figure 19). Used with permission of American Psychological Association, Inc.	183
6.1	A numeral factorization transducer for numbers up to 999.	201
6.2	Two-dimensional layout in an e-mail signature.	212

List of Tables

2.1	Chinese characters illustrating the five modes of combination of semantic (underlined) and phonetic components.	35
2.2	Full and diacritic forms for Devanagari vowels, classified by the position of expression of the diacritic forms. Thus “after” means that the diacritic occurs after the consonant cluster, “below,” below it, and so forth.	46
2.3	Distribution of placement of semantic component (Kang-Xi Radical) among 12,728 characters from the Taiwan Big5 character set.	51
2.4	Distribution of placement of semantic component (Kang-Xi Radical) among 2,588 characters from the Taiwan Big5 character set.	52
2.5	Linguistic units corresponding to Mayan complex glyphs, from Macri (1996); glossing conventions follow those of Macri. In some phrasal cases it is unclear whether the unit in question is really a constituent: Such cases are indicated with a question mark.	56
3.1	The ten most frequent lexical markings for the shallow ORL in the English fragment.	87
3.2	The ten most frequent lexical markings for the deep ORL in the English fragment.	87
4.1	Disyllabic morphemes collected from the ROCLING corpus (10 million characters) and 10 million characters of the United Informatics corpus. This set consists of pairs of characters occurring at least twice, and where each member of the pair only cooccurs with the other.	149
4.2	Further disyllabic morphemes collected from the ROCLING corpus (10 million characters) and 10 million characters of the United Informatics corpus. This set consists of other pairs of characters that do not exclusively occur with each other, but where there is nonetheless a high mutual information for the pair. Note that LV (†) indicates that the phonetic component in question occurs 9	

- out of 38 times in characters pronounced with initial /l/
followed by some vowel.

4.3 A sample of Japanese kokuji (second column), with their
componential analysis (third column). The first column is
the entry number in Alexander's list (Lehman and Faust,
1951). The fourth column lists the phonetic, if any. The
fifth column lists the kun pronunciation, and the sixth
column the *on* pronunciation, if any: In one case – *goza* –
there is no kun pronunciation. The last three kokuji shown
are formed as semantic–phonetic constructs, with the last
two being based on the kun pronunciation: Note that the
phonetic component of *masa* also means 'straight', so it is
possible that this one is also a semantic–semantic construct.

6.1 Total counts in Cregeen's dictionary of words spelled with
initial <Ch>, where *C* denotes a consonant, and the
number and percentages of those words that are minimal
pairs with homophonic or close-homophonic words spelled
without the <h>.

150

156

191

1 Reading Devices

Our starting point for this study of writing systems is text-to-speech synthesis – TTS, and more specifically the computational problem of converting from written text into a linguistic representation. While the connection between TTS systems on the one hand and writing systems on the other may not be immediately apparent, a moment's reflection will make it clear that the problem to be solved by a TTS system – namely the conversion of written text into speech – is exactly the same problem as a human reader must solve when presented with a text to be read aloud. And just as writing systems, their properties, and the ways in which they encode linguistic information are of interest to psycholinguists who study how people read, so (in principle) should such considerations be of interest to those who develop TTS technology: At the very least, it ought to be of as much interest as, for example, understanding the physiology and acoustics underlying speech production, something that early speech synthesis researchers such as Fant (1960) were heavily involved in.¹

Since my starting point is TTS, and since I assume that most readers will not be familiar with this field, I will start this chapter with a review of some of the issues relevant to the development of TTS systems, particularly as they relate to the problem of analyzing input text. This will be the topic of Section 1.1. In Section 1.2 I will informally introduce, by way of a simple example, the model that I shall be developing throughout the rest of this book. Finally, Section 1.3 will introduce some aspects of the formalism and the conventions that will be used throughout this book.

¹ It will perhaps come as no surprise that TTS researchers have *not*, in fact, generally been overly interested in writing systems. This is undoubtedly due in part to the relatively low interest in text-analysis issues in general in the TTS literature, at least as compared to the high level of interest in such matters as prosody, intonation, voice quality, and synthesis techniques. It also is undoubtedly related to the fact that much of the work on TTS is driven by rather practical aims (e.g., building a working system), where an overactive interest in theories of writing systems might appear to be an unnecessary luxury.

1.1 Text-to-Speech Conversion: A Brief Introduction

As noted above, the task of a TTS system is to convert written text into speech. Normally the written representation is in the form of an electronic text – coded in ASCII, ISO, JIS, UNICODE, or some other standard depending upon the language and system being used; this circumvents one problem that humans must solve, namely that of visually recognizing characters printed on a page.² Similarly the output is a digital representation of speech. Between these two representations are numerous stages of processing, which can be profitably classified into two broad stages. The first stage is the conversion of the written text into an internal linguistic representation; the second is the conversion from that linguistic representation into speech. The latter consists of computing various phonetic and acoustic parameters, including segmental duration, F_0 (“pitch”) trajectory, properties of the output speech such as spectral tilt or glottal open quotient, and (in concatenative speech synthesis systems) selection of appropriate acoustic units or (in formant-based synthesis systems) the generation of vocal-tract transfer functions appropriate to the intended sounds. We will have nothing further to say about these issues here; the reader is referred to Dutoit (1997) for a good general introduction to these issues and also to Allen, Hunnicutt, and Klatt (1987) and Sproat (1997b) for an overview of how two particular systems (the MITalk system and the Bell Labs TTS system) work.

In any TTS system the output speech will be generated from an annotated linguistic representation, which is in turn derived from input text via the first stage of processing defined above. How rich a linguistic representation is presumed (and in terms of which linguistic theories and assumptions it is couched) differs from system to system, of course, but we may at least assume that the linguistic representation will include information on the sequence of sounds to be enunciated (usually allophones of phonemes, but in some systems whole syllable-sized units); lexical stress or tone information; word and phrase-level accentuation and emphasis; and the location of various prosodic boundaries, including syllable and prosodic phrase boundaries. Thus for an input such as that in (1.1), we might presume as a plausible (partial) linguistic representation, the representation in Figure 1.1.

(1.1) I need 2 oz. of Valrhona and 6 anchos for the mole.

In the particular rendition of the sentence presumed in Figure 1.1 there are two intonational phrases (denoted by ι), grouped into a single utterance (U). Lexical stress is indicated by a metrical tree dominating individual

² Of course, it is possible to hook up a TTS system to an *optical character recognition* (OCR) system; such systems have in fact been available for several years in the form of page-readers for the blind (e.g., Kurzweil’s reader); and there has been much recent interest in conversion of FAX into speech, which adds yet a further complication, namely messy input.

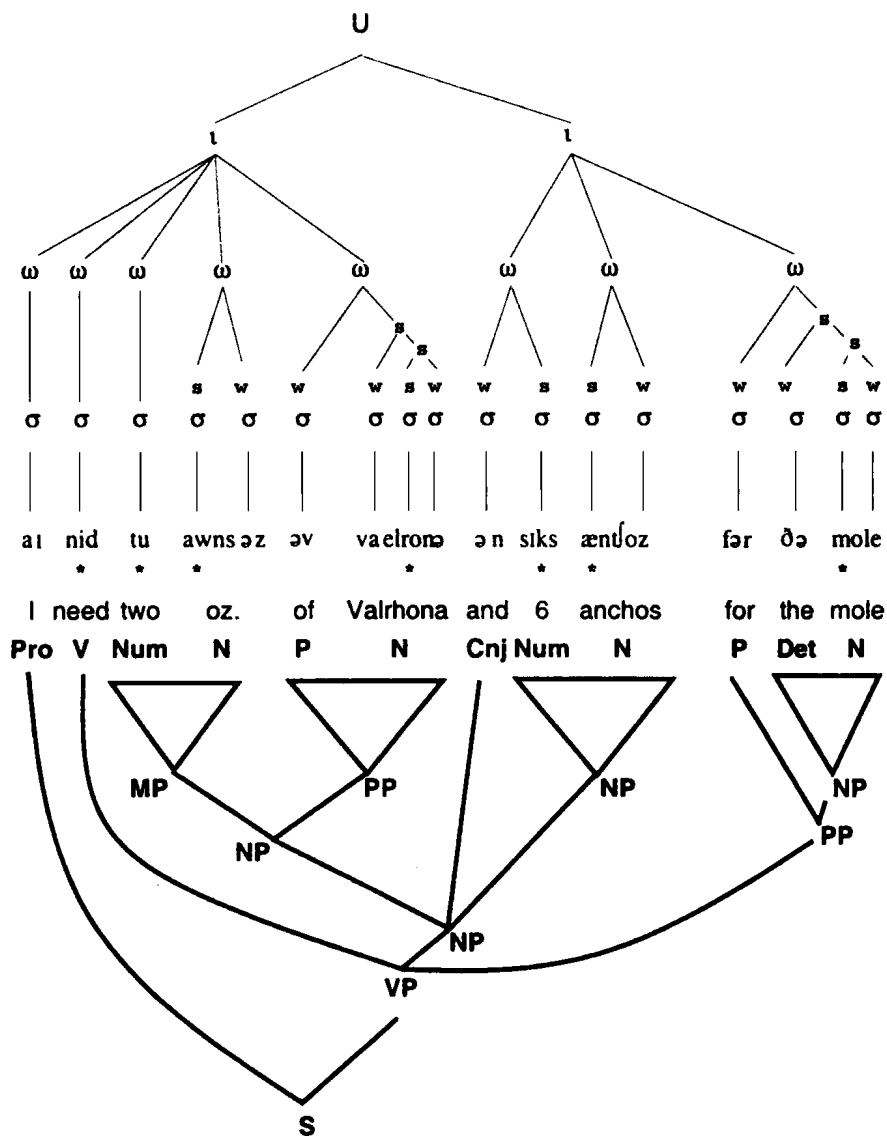


Figure 1.1. A partial linguistic representation for the sentence in (1.1). Shown are a phonetic transcription, a prosodic analysis into two intonational phrases (ι) and one utterance (U), accent assignment (*), a set of part of speech tags, and a simple phrase-structure analysis. Phonetic symbols are IPA. Note that 'MP' means 'measure phrase'.

syllables (σ) and dominated itself by a prosodic word (ω); we assume that proclitics form a prosodic word with the following content word. Also indicated are lexical accents for the words *need*, *two*, *ounces*, *Valrhona*, *six*, *anchos*, and *mole*.

To produce this representation, or any equally plausible representation, for this sentence, a reader must “reconstruct” a great deal of linguistic information that is simply not represented in the written form. Naturally all syntactic information, including both the morphosyntactic part of speech tags as well as phrase structure, must be computed. So must a great deal of the phonological information. In particular, the sequence of phonetic segments are only somewhat indirectly represented in English orthography; in some written forms such as *2*, *6*, and *oz.* they cannot be said to be represented at all. In the latter case the linguistic form must be reconstructed entirely from the reader's knowledge of the language and often depends upon information about context (does one say *ounce* or *ounces*?). In some cases readers may need to make educated guesses about the pronunciations of some words, though if these follow the normal pronunciation conventions of the language they will usually guess correctly: Even readers who had not previously seen the words *anchos* or *Valrhona* could nonetheless probably have guessed the correct pronunciation. For *mole* – in the sense of a Mexican sauce, and pronounced /*molei*/ – the situation is more complex since the pronunciation here does not follow standard English conventions: In this case one would simply have to be familiar with the word. But there is of course an additional problem here in that, as in the case of *oz.*, one must also disambiguate this word, so that one does not pronounce it as the homographic /*mol*/ (e.g., in the sense of a species of insectivore).

Prosodic phrasing is rarely represented; note that punctuation is only partly used in this function (Nunberg, 1995), and in any case it is by no means consistently used in every case where one might plausibly find a prosodic boundary. Lexical accentuation is almost never indicated.³

Thus, if one is designing a TTS system that can handle arbitrary text in a given language, it is generally necessary for the system to possess a large

³ It is generally true that suprasegmental and prosodic information is systematically omitted from the orthographies of a large variety of languages. This is particularly true for high level prosodic information such as prosodic phrase boundary placement, and accentuation and prominence. But it extends to purely lexically determined features such as lexical tone. Thus while some languages, such as Thai, Vietnamese, or Navajo, *do* indicate lexically distinctive tone in their orthographies, it seems to be far more common to omit this feature: For example many orthographies developed for tonal languages of Africa omit marks of tone, though it should be noted that many of these scripts were developed by European missionaries who had no understanding of tone; see Bird (1999) for a discussion of more recently developed African orthographies where tone is marked.

A related point, as Geoffrey Sampson has noted (personal communication), is that Latin did not mark length in vowels (though gemination in consonants was marked).