DESERET LANGUAGE

AND

LINGUISTIC SOCIETY

SELECTED PAPERS FROM THE

PROCEEDINGS

FIFTEENTH ANNUAL SYMPOSIUM

13-14 MARCH 1989

BRIGHAM YOUNG UNIVERSITY
PROVO, UTAH

EDITED BY:
SOREN F. COX
ENGLISH DEPARTMENT
BRIGHAM YOUNG UNIVERSITY

COPYRIGHT 1989 DLLS
SELECTED PAPERS FROM THE
PROCEEDINGS
FIFTEENTH ANNUAL SYMPOSIUM
13-14 MARCH 1989
BRIGHAM YOUNG UNIVERSITY
PROVO, UTAH
EDITED BY:
SOREN F. COX
ENGLISH DEPARTMENT
BRIGHAM YOUNG UNIVERSITY
COPYRIGHT 1989 DLLS
ACKNOWLEDGEMENTS

The Deseret Language and Linguistics Society expresses thanks for the generous support of its Annual Symposium by the Brigham Young University College of Humanities. Thanks also are extended to the Linguistics Department and the English Department for their assistance. We are grateful to those who took part in any way in organizing the program, chairing the sessions, presenting papers, or attending the symposium.

The symposium program with names of authors and titles of papers is given on the following pages. The Table of Contents contains the names of the authors and the titles of papers submitted for publication in the proceedings. Individual authors are responsible for the content and accuracy of their respective papers.

Soren F. Cox, President
Deseret Language and Linguistics Society
### Monday, March 13, 1989

<table>
<thead>
<tr>
<th>Presenter(s)</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alan C. Covell</td>
<td>&quot;For Want of a Nail&quot;: Should We Teach Pronunciation to Teach Grammar?</td>
</tr>
<tr>
<td>Kenneth R. Beesley</td>
<td>Computer Analysis of Aymara Morphology</td>
</tr>
<tr>
<td>Stuart N. Newton</td>
<td></td>
</tr>
<tr>
<td>Joni M. Kroepelin</td>
<td>Is He Considered a Neutral Pronoun?</td>
</tr>
<tr>
<td>Marvin Folsom</td>
<td>The Language of Orson Hyde's <em>Ein Ruf aus der Wuste</em></td>
</tr>
<tr>
<td>Randall Jones</td>
<td>The Preparation of Texts for Computer Processing</td>
</tr>
<tr>
<td>Kim Clausson</td>
<td>The Sacred and the Holy in the Book of Mormon</td>
</tr>
<tr>
<td>Earlene Sudweeks</td>
<td>Computer Aided Spanish Application</td>
</tr>
<tr>
<td>Barbara Gordon</td>
<td></td>
</tr>
<tr>
<td>Royal Skousen</td>
<td>Early Textual Errors in the Book of Mormon</td>
</tr>
<tr>
<td>Larry E. Wood</td>
<td>Natural Language Processing Applied to a Database of Engineering Design Documents</td>
</tr>
<tr>
<td>Randy L. Orr</td>
<td></td>
</tr>
<tr>
<td>Alan R. Parkinson</td>
<td></td>
</tr>
<tr>
<td>Keynote Address:</td>
<td>Questions of Adult Language Learnability</td>
</tr>
<tr>
<td>William Rutherford</td>
<td></td>
</tr>
<tr>
<td>Alan Melby</td>
<td>Electronic Bilingual Dictionaries and the Language Lab</td>
</tr>
<tr>
<td>Don Chapman</td>
<td>Politics and Orwell's Language</td>
</tr>
<tr>
<td>Larry G. Childs</td>
<td>Machine Translation and the Word Processor</td>
</tr>
<tr>
<td>Kenton L. Chauncey</td>
<td>Automatic Recognition of Relative Clauses with Missing Relative Pronoun</td>
</tr>
<tr>
<td>Marian McMaster</td>
<td>What Every Would-Be Translator Should Know about Translation</td>
</tr>
<tr>
<td>Mike L. Reynolds</td>
<td>Metatext: Computerized Materials for the Study of Shakespeare's Language</td>
</tr>
<tr>
<td>Camille Williams</td>
<td></td>
</tr>
</tbody>
</table>
Tuesday, March 14, 1989

William Eggington
- Topic Development in Australian English Texts

Deryle Lonsdale
- Reaccenting French Text

Robert L. Good
- Supermarket Lexicography or Finding New Terms to Put in the Dictionary

Russell Bateman
- Syntactic Complexity in Virgil's Aeneid

Melvin Luthy
- A Computerized Rhyming Dictionary/Research Tool

Robert Stevens

Danial Parvaz

Wendy B. Taylor
- Yiddish and ASL: A Look at Linguistic Prejudice

Chauncey Riddle
- Teaching A Man to Fish

Keynote Speaker Seminar:
William Rutherford
- Truth and Language
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim Clausson</td>
<td>The Sacred and the Holy in the Book of Mormon</td>
<td>1</td>
</tr>
<tr>
<td>Marvin Folsom</td>
<td>The Language of Orson Hyde's Ein Ruf aus der Wuste</td>
<td>11</td>
</tr>
<tr>
<td>Chauncey Riddle</td>
<td>Truth and Language</td>
<td>19</td>
</tr>
<tr>
<td>Mike L. Reynolds and Camille S. Williams</td>
<td>Metatext: Computerized Materials for the Study of Shakespeare's Language</td>
<td>26</td>
</tr>
<tr>
<td>Deryle Lonsdale</td>
<td>The Reaccentuation of French Text</td>
<td>41</td>
</tr>
<tr>
<td>Don Chapman</td>
<td>Orwell’s Language and Thought in &quot;Politics and the English Language&quot; and 1984</td>
<td>55</td>
</tr>
<tr>
<td>Marian McMaster-Labrum</td>
<td>What Every Would-Be Translator Should Know about Translation</td>
<td>63</td>
</tr>
<tr>
<td>William Eggington</td>
<td>Topic Development in Australian Aboriginal English Texts</td>
<td>68</td>
</tr>
<tr>
<td>Joni M. Kroepelin</td>
<td>Is He Considered a Gender Neutral Pronoun?</td>
<td>86</td>
</tr>
<tr>
<td>Robert L. Good</td>
<td>Supermarket Lexicography or Finding New Terms to Put in the Dictionary</td>
<td>104</td>
</tr>
<tr>
<td>Alan Melby</td>
<td>Electronic Bilingual Dictionaries in the Language Lab</td>
<td>111</td>
</tr>
<tr>
<td>Melvin Luthy and Robert Stevens</td>
<td>A Computer Rhyming Dictionary/Research Tool</td>
<td>112</td>
</tr>
<tr>
<td>Kenton L. Chauncey</td>
<td>Automatic Recognition of Relative Clauses with Missing Relative Pronoun</td>
<td>117</td>
</tr>
<tr>
<td>Russell Bateman</td>
<td>Syntactic Complexity in Virgil’s Aeneid</td>
<td>145</td>
</tr>
<tr>
<td>Larry E. Wood, Randy L. Orr, and Alan R. Parkinson</td>
<td>Natural Language Processing Applied to a Database of Engineering Design Documents</td>
<td>151</td>
</tr>
</tbody>
</table>
Larry G. Childs
For Want of a Nail: Machine Translation and the Word Processor 160

Earlene Sudweeks and Barbara Gordon
Computer Assisted Spanish Applications (CASA) 168

Danial Parvaz
Yiddish and ASL: A Look at Linguistic Prejudice 171

Wendy B. Taylor
Teaching A Man to Fish 177

Alan C. Covell
For Want of a Nail: Should We Teach Pronunciation to Teach Grammar? 184

Royal Skousen
Progress Report on the Book of Mormon Critical Text Project 194
One day I looked up the word 'sacred' in a dictionary. The definition it gave was "holy." When I looked up 'holy,' that was defined as "sacred." My dissatisfaction motivated an experiment. I went through the Book of Mormon and made two lists. One was a list of each instance of the word 'sacred' and what it modified or referred to. Then I did the same for the word 'holy.' When I looked over my lists, I was surprised to see that they were mutually exclusive. In the Book of Mormon only a few specific things are called "sacred," while many other things are called "holy."1

By contrast, the word 'sacred' does not appear at all in the King James Version of the Holy Bible.2 The King James Version was published in England in 1611. The Book of Mormon was translated from ancient metal plates by Joseph Smith in Pennsylvania and New York in 1828 and 1829 and published in 1830.

The word 'sacred' is used in the Book of Mormon primarily in the context of engraving the plates with the scriptures and transferring the plates and their associated revelatory instruments, the pair of stones called the interpreters3 and the ball or director called the Liahona4, from a prophet to his successor.

The first appearance of the term is in the First Book of Nephi. Nephi explains that he was commanded to "make plates of ore that I might engraven upon them the record of my people." 1 Ne.19:1. Later he is commanded to make other plates, the "small plates," "that the ministry and the prophecies, the more plain and precious parts of them, should be written upon these plates; and that the things which were written should be kept for the instruction of my people ... and also for other wise purposes, which purposes are known unto the Lord." 1 Ne.19:3,5

Nephi continues, "I proceed according to that which I have spoken: and this I do that the more sacred things may be kept for the knowledge of my people. Nevertheless, I do not write anything upon plates save it be that I think it be sacred." 1 Ne.19:5-6.

After Nephi's death, his brother Jacob writes that Nephi instructed him "that I should preserve these plates and hand them down unto my seed, from generation to generation. And if there was preaching which was sacred, or revelation which was great, or prophesying, that I should engraven the heads of them upon these plates." Jacob 1:4.
The next instance of the word 'sacred' is also connected with the transfer of the plates from a prophet to his successor. Alma "the younger" tells his son Helaman to

take the records which have been entrusted with me; And I also command you that ye keep a record of this people, according as I have done, upon the plates of Nephi, and keep all these things sacred which I have kept, even as I have kept them; for it is for a wise purpose that they are kept. Alma 37:1-2.

In the same conversation Alma explains, "remember, my son, that God has entrusted you with these things, which are sacred, which he has kept sacred, and also which he will keep and preserve for a wise purpose in him, that he may show forth his power unto future generations." Alma 37:14. Alma warns Helaman that "if ye transgress the commandments of God, behold, these things which are sacred, shall be taken away from you by the power of God." Alma 37:15. On the other hand,

if ye keep the commandments of God, and do with these things which are sacred according to that which the Lord doth command you, (for you must appeal unto the Lord for all things whatsoever ye must do with them) behold, no power of earth or hell can take them from you, for God is powerful to the fulfilling of all his words. Alma 37:16

After discoursing on the interpreters and the Liahona, which he also transferred to Helaman, Alma entreats. "And now, my son, see that ye take care of these sacred things." Alma 37:47.

In one of the prophet Mormon's references to the transfer between Alma and Helaman, he points out that Nephihah, the second chief judge, "had refused Alma to take possession of those records and those things which were esteemed by Alma and his fathers to be most sacred; therefore Alma conferred them upon his son, Helaman." Alma 50:38.

Later Mormon records that "Shiblon took possession of those sacred things which had been delivered unto Helaman by Alma." Alma 63:1. Three years later "it became expedient for Shiblon to confer those sacred things, before his death, upon the son of Helaman, who was [also] called Helaman." Alma 63:11.

[All those engravings which were in the possession of Helaman were written and sent forth among the children of men throughout all the land, save it were those parts which had been commanded by Alma should not go forth. Nevertheless, these things were to be kept sacred, and handed down from one generation to another. Alma 63:12-13.
Fifty years later

Nephi, the son of Helaman, had departed out of the land of Zarahemla, giving charge unto his son Nephi, who was his eldest son, concerning the plates of brass, and all the records which had been kept, and all those things which had been kept sacred from the departure of Lehi out of Jerusalem. 3 Ne.1:2

The prophet Mormon describes his own predecessor's actions in these words: "Ammaron, being constrained by the Holy Ghost, did hide up the records which were sacred -- yea, even all the sacred records which had been handed down from generation to generation, which were sacred." 4 Ne.1:48. When Mormon was ten years old Ammaron told him, "when ye are about twenty and four years old ... go to the land Antum, unto a hill which shall be called Shim; and there have I deposited unto the Lord all the sacred engravings concerning this people." Mormon 1:3.

The word 'sacred' is used again when Mormon describes his own transfer of the plates to his son Moroni. He writes,

I, Mormon, began to be old; ... and having been commanded of the Lord that I should not suffer the records which had been handed down by our fathers, which were sacred, to fall into the hands of the Lamanites, (for the Lamanites would destroy them) therefore I made this record out of the plates of Nephi, and hid up in the hill Cumorah all the records which had been entrusted to me by the hand of the Lord, save it were these few plates which I have unto my son Moroni. Mormon 6:6

Mormon writes to Moroni that "I have sacred records that I would deliver up unto thee." Moroni 9:24

Moroni, in abridging the Jaredite plates, refers to a hill named Ramah as "that same hill where my father Mormon did hide up the records unto the Lord, which were sacred." Ether 15:11.

The word 'sacred' in Helaman 4:12 is not unambiguously used in the context of the plates, but it may refer to them. In that verse, in a series of parallel participial phrases, Mormon lists the character defects of the Nephites, which include "making a mock of that which was sacred, denying the spirit of prophecy and of revelation." Prophecy and revelation are connected with the word 'sacred' and with the plates in Nephi's instruction to Jacob that "if there were preaching which was sacred, or revelation which was great, or prophesying," he was to "engraven the heads of them upon these plates." Jacob 1:4.

We now consider the remaining few instances of the word 'sacred' which occur in contexts other than those involving the plates.
At the end of a bloody battle Moroni, the chief commander over the armies of the Nephites, shouts to his opponent.

Zarahemnah, I command you, in the name of that all-powerful God who has strengthened our arms that we have gained power over you, by our faith, by our religion, and by our rites of worship, and by our church, and by the sacred support which we owe to our wives and our children, by that liberty which binds us to our lands and our country; yea, and also by the maintenance of the sacred word of God, to which we owe all our happiness; and by all that is most dear unto us -- Yea, and this is not all; I command you by all the desires which ye have for life, that ye deliver up your weapons of war unto us, and we will seek not your blood, but we will spare your lives, if ye will go your way and come not again to war against us. Alma 44:5-6.

The "sacred word of God" is engraved upon the plates.

Mormon writes that Pahoran "was appointed chief judge and governor over the people, with an oath and sacred ordinance to judge righteously, and to keep the peace and the freedom of the people, and to grant unto them their sacred privileges to worship the Lord their God." Alma 50:39.7

In another list of condemned practices, Mormon records that at the beginning of the third century after Christ's visit to the American continent, "many churches ... did administer that which was sacred unto him to whom it had been forbidden because of unworthiness." 4 Ne.1:27. This apparently is a reference to the ordinance of the sacrament of the Lord's supper involving the ritual partaking of sanctified bread and wine.8

In the Book of Mormon there are 27 instances of the word 'sacred.' Twenty-one of those instances (78%) are in the context of the plates and associated revelatory instruments, 1 instance modifies "word of God," and 1 appears to refer to "prophecy and revelation," both of which are metonymically related to the plates. Two of the remaining instances of the word 'sacred' refer to ordinances: the ordinance of appointment of the chief judge and the ordinance of the sacrament of the Lord's supper. The reference to "sacred privileges to worship" may be related to ordinances, especially considering Moroni's reference to "our rites of worship" in the same speech. Finally, there is the reference to the support of wives and children.

Note the kinds of things described as "sacred": the word of God, plates containing the word of God, preaching, prophesying, revelations, revelatory instruments, ordinances, privileges, and obligations. Not persons. Not *sacred prophets, *sacred men, or *sacred God. Not places. Not *sacred city, or *sacred hill.
In contrast to 'sacred,' the word 'holy' is used to describe persons, places, and a wide variety of things, but not the same things as 'sacred.' There are 297 instances of 'holy' in the Book of Mormon, 11 times as many as 'sacred.' The following table outlines the instances of 'holy' in the Book of Mormon.

<table>
<thead>
<tr>
<th>Word Modified</th>
<th>Number of Instances</th>
<th>% of Total Instances</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ghost</td>
<td>94</td>
<td>32</td>
<td>38 on small plates, 1 Jesus quoting Isaiah in 3 Ne. 22:5</td>
</tr>
<tr>
<td>One of Israel</td>
<td>39</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>prophet(s)</td>
<td>28</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>order</td>
<td>17</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Spirit</td>
<td>16</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>One</td>
<td>9</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>God</td>
<td>8</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>calling</td>
<td>7</td>
<td>2</td>
<td>none on small plates</td>
</tr>
<tr>
<td>man/men</td>
<td>7</td>
<td>2</td>
<td>none on small plates</td>
</tr>
<tr>
<td>name</td>
<td>7</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>city</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>arm</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>scriptures</td>
<td>5</td>
<td>2</td>
<td>all in Alma</td>
</tr>
<tr>
<td>angels</td>
<td>4</td>
<td>1</td>
<td>none on small plates</td>
</tr>
<tr>
<td>judgment(s) of God</td>
<td>3</td>
<td>1</td>
<td>all on small plates</td>
</tr>
<tr>
<td>Lamb of God</td>
<td>3</td>
<td>1</td>
<td>all on small plates</td>
</tr>
<tr>
<td>Lord of Hosts</td>
<td>3</td>
<td>1</td>
<td>all on small plates</td>
</tr>
<tr>
<td>sabbath day</td>
<td>3</td>
<td>1</td>
<td>all on small plates</td>
</tr>
<tr>
<td>ye</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Messiah</td>
<td>2</td>
<td>2</td>
<td>both in 2 Ne. 2</td>
</tr>
<tr>
<td>mountain</td>
<td>2</td>
<td>2</td>
<td>none on small plates</td>
</tr>
<tr>
<td>ones</td>
<td>2</td>
<td>2</td>
<td>both in Alma</td>
</tr>
<tr>
<td>stand</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thou (God, Lord)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>work(s)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Being</td>
<td>1</td>
<td></td>
<td>Mormon</td>
</tr>
<tr>
<td>Child</td>
<td>1</td>
<td></td>
<td>Moroni</td>
</tr>
<tr>
<td>children</td>
<td>1</td>
<td></td>
<td>Alma</td>
</tr>
<tr>
<td>church</td>
<td>1</td>
<td></td>
<td>Mormon</td>
</tr>
<tr>
<td>commandments</td>
<td>1</td>
<td></td>
<td>Alma</td>
</tr>
<tr>
<td>fathers</td>
<td>1</td>
<td></td>
<td>Alma</td>
</tr>
<tr>
<td>land</td>
<td>1</td>
<td></td>
<td>small plates</td>
</tr>
<tr>
<td>One of Jacob</td>
<td>1</td>
<td></td>
<td>small plates</td>
</tr>
<tr>
<td>ordinance</td>
<td>1</td>
<td></td>
<td>Alma</td>
</tr>
<tr>
<td>people</td>
<td>1</td>
<td></td>
<td>Alma</td>
</tr>
<tr>
<td>sanctuary</td>
<td>1</td>
<td></td>
<td>Ether</td>
</tr>
<tr>
<td>seed</td>
<td>1</td>
<td></td>
<td>2 Ne.</td>
</tr>
<tr>
<td>that which is</td>
<td>1</td>
<td>3 Ne.</td>
<td></td>
</tr>
<tr>
<td>they (Israelites)</td>
<td>1</td>
<td>small plates</td>
<td></td>
</tr>
<tr>
<td>they (3 Nephite</td>
<td>1</td>
<td>3 Ne.</td>
<td></td>
</tr>
<tr>
<td>disciples)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>will</td>
<td>1</td>
<td>Moroni</td>
<td></td>
</tr>
<tr>
<td>word of God</td>
<td>1</td>
<td>Mormon</td>
<td></td>
</tr>
<tr>
<td>writ</td>
<td>1</td>
<td>Alma</td>
<td></td>
</tr>
</tbody>
</table>
Grammatical Observations

'Sacred' and 'holy' are used as adjectives in three syntactic constructions.

[article] + 'sacred' + [noun]
'those'

the sacred records
the more sacred things
the holy prophets
a holy man

[noun] + 'be' + 'sacred'

things which are sacred
records ... esteemed to be most sacred
thou art holy

'keep' + [noun] + 'sacred' OR [noun] + 'kept' + 'sacred'

keep all these things sacred
which he has kept sacred
observe the sabbath day, and keep it holy

Semantic Observations

The differences in the uses of 'sacred' and 'holy' in the Book of Mormon suggest connotations which we may not feel in standard English today or even in current LDS usage.

'Holy' suggests well-being, integrity, blessedness, as a sign of divine grace. A holy person is one blessed by God, perhaps even "charged with the divine presence."11 This word has a much broader range of application than does 'sacred' and is particularly (perhaps made more apparent in contrast to 'sacred') appropriate as a personal epithet.

'Sacred' has an ambivalent character which suggests consecrated to God and associated with terrible penalties, "august and accursed, worthy of veneration and evoking" fear.12 The things referred to as "sacred" in the Book of Mormon have awful qualities as well as sublime ones. Alma's warnings to his son Helaman about his obligations while in possession of the plates and related things and the awful consequences of his disobedience sound a lot like Moroni's warnings to Joseph Smith when he was given the plates and related things. Both were told that they must keep the plates and related items away from the world, they must "keep them sacred," so that "the sacred word of God" written upon them could be preserved for God's purposes.13
The sacred revelatory instruments included the interpreters and the Liahona. Like the plates, both of these items served as transmitters of the word of God. Perhaps these could be characterized as dynamic transmitters of God's word as opposed to the static word of God engraved upon the plates. Both sets of transmitters have a direct, close connection with God, are vehicles of direct communication from God. They are sacred because they transmit "that which cometh from above," which is sacred. And they belong to God. They are not of this world in the sense of being in regular circulation in the world.

The sacred ordinance, their sacred privileges to worship God, the sacred sacrament, the sacred spirit of prophecy and of revelation are all things God gave to the people to put them in contact with God. The sacred support which they owed to their wives and their children was an opportunity with an obligation put on them by God with awful consequences for its disregard.

'Sacred' is not applied to persons in the Book of Mormon. God's prophets sent to the people to speak the word of God are never called "sacred." Holy yes. Not sacred. Sacred when applied to a person would probably mean dead. Cut off from this world. The victim in the sacrifice is made sacred by being cut off from this world to cross the veil into the other world and make contact with God on behalf of the worshippers or devotees that sacrificed it.

On a rather abstract level, one dimension of the differences between 'sacred' and 'holy' in the Book of Mormon can be described as 'sacred' involves contiguity to God, while 'holy' suggests similarity to God. 'Sacred' is metonymic, 'holy' is metaphoric.

Conclusion

'Sacred' and 'holy' are used in markedly different contexts in the Book of Mormon. Three-fourths of the instances of 'sacred' occur in the context of the plates, while the rest refer to non-personal entities like ordinances, privileges, and obligations that constitute connections between people and God. In the Book of Mormon 'sacred' suggests contiguity with God, radical separation from this world. 'Holy' occurs 11 times as often as 'sacred' and is used to describe persons, places, and things that enjoy or manifest God's grace or blessing.

A somewhat similar pattern exists in the Doctrine and Covenants, where 16 of the 21 instances of 'sacred' (76%) are in the context of plates or the translation or publication of the scriptures. 'Holy' occurs 163 times in the Doctrine and Covenants, 7 times as often as 'sacred.' They are joined in "sacred and holy purposes" and in "holy and sacred writings." D&C 104:65, 68. 'Sacred' does not occur in the Pearl of Great Price.
A preliminary study of Joseph Smith's noncanonical writings and orations shows that he used 'holy' about 9 times as often as 'sacred.' Only about a quarter of the instances of 'sacred' were in the context of the scriptures. His uses of 'sacred' often describe remembrance, burial, and blood of Revolutionary heroes. Joseph Smith's uses of 'holy' and 'sacred' are certainly a fruitful object of further study.

Another interesting subject for study is contemporary LDS usage of 'holy' and 'sacred.' While I have not begun to even preliminarily study this phenomenon, my own surprise at the Book of Mormon pattern of usage of 'holy' and 'sacred' suggests that contemporary LDS usage is quite different from that found in the Book of Mormon, where sacred is more sacred than holy.

Footnotes

1Shapiro 1973:453-55,822-23. 'Profane' is only in Jarom 1:5: '[T]hey profaned not; neither did they blaspheme." Four of the 5 instances of 'unholy' are in the expression "dwelleth not in unholy temples." Mos.2:37; Alma 7:21, 34:36; Hel. 4:24. One other contrasts "we, being unholy," with "the Lamb of God, he being holy." 2 Ne. 31:5. 'Unholy' is in the Doctrine and Covenants (hereafter "D&C") only thrice, 74:4,6, 97:17, and 'profane' not at all. Shapiro 1973:767,1002. 'Unholy' and 'profane' are rare in Joseph Smith's noncanonical writings. He used 'profane' only to describe language. Madsen 1985:318,426.

2Cruden 1953; Strong 1890. But see Gileadi 1982:145 (Isaiah 57:15) where the name of the Lord is "sacred." 'Sacred' is in the Douay Bible, published about the same time as King James. Thompson & Stock 1945. Ex. 13:5, 31:10, 39:29, & 2 Ma 4:48 are translations of Vulgate 'sacer' and Num 27:11 of 'sanctus'.

3Also called "directors" in Alma 37:21,24 in original and printer's manuscripts and in published editions of the Book of Mormon from 1830 through 1911, changed to "interpreters" in 1920 and 1981. FARMS 1986:674,n.622,675,n.625. Also called "Urim and Thummim" by Joseph Smith, e.g., Pearl of Great Price (hereafter "POGP"), JS-H 1:35,42,52,59,62.

4See 1 Ne.16:10,26-30; Alma 37:38. Also called a "compass" in 1 Ne.18:12,21; 2 Ne.5:12.

5See also Words of Mormon 1:3-11; D&C 10:9("sacred"),38-45.

6Compare the warnings of Moroni to Joseph Smith concerning his possession and use of the plates and interpreters, e.g., in POGP, JS-H 1:42,46,59-60.

7Compare D&C 134:5 in connection with 134:2,4,6 & POGP, Article of Faith 11.
9The one thing that is called both "holy" and "sacred."


12See Benveniste 1973:452. For an interesting study of 'holy,' 'sacer,' 'sanctus,' and other items of Indo-European religious vocabulary, see Benveniste 1973:445-528.

13Alma 37:2, 44:5; POGP, JS-H 1:42,46,59-60. Although 'sacred' is not found in the JS-H 1 description of his conversations with Moroni, Joseph Smith did use that word in a diary entry on the subject: "he told me of a sacred record which was written on plates of gold." Jessee 1984:76.

14D&C 6:10, 63:64.


17Shapiro 1977:456,823. The portions of D&C & POGP considered in this study record revelations and writings of Joseph Smith between 1828 and 1843.

18Madsen 1985:192-93,357.


Bibliography


*Book of Mormon*. 1981. Church of Jesus Christ of Latter-day Saints, Salt Lake City.


*Doctrine and Covenants of the Church of Jesus Christ of Latter-day Saints* [and] *Pearl of Great Price*. 1982. Church of Jesus Christ of Latter-day Saints, Salt Lake City.


*Holy Bible* (King James). 1979. Church of Jesus Christ of Latter-day Saints, Salt Lake City.


Strong, J. 1890. *The Exhaustive Concordance of the Bible*. Abingdon, N.Y.

In the April 6th conference of 1840, Orson Hyde referred to prophecy concerning the great work to be performed among the Jews. On the 15th of April (1840) he left Commerce for Jerusalem. In a letter to Joseph Smith dated May 1, 1840, he discussed his plan to write a set of lectures upon the faith and doctrine of the church, giving a brief history of the coming forth of the Book of Mormon and an account of its contents, together with the outlines of the organization and government of the church. He wanted to have it translated into German, have it published when he arrived in Germany and then distribute it throughout the German empire. In January 15, 1841, Orson Hyde and John Page were informed that the Lord was displeased with them for delaying their mission and were instructed to hasten their journey towards their destination. Orson Hyde set sail from New York for Liverpool on February 13, 1841. From a letter to Joseph Smith, we know that he completed the book by June 15, 1841. He left London for Rotterdam on June 20. In July, he was in Regensburg waiting to get a visa and decided to spend his time studying German. He became acquainted with a lady who spoke French and German and who was anxious to speak English. He gave her English lessons and she gave him German lessons. On Sunday, October 24, 1841 he dedicated the Holy Land. In November, he was in Alexandria and in January in Triest. On November 22, 1841, he wrote to Joseph Smith telling him of his plans to go to Regensburg where he arrived in February of 1842 and where some friends were willing to help him "publish our faith in the German language." With the help of one of his students to whom he was giving English lessons, he completed the translation of the book (of 115 pages) and submitted it to the censors, who did not approve it for publication. He then went to Frankfurt where he received permission to publish it sometime later in 1842, as the title page indicates Im Selbstverlage des Verfassers.

I have been working on a translation of the section in the book that deals with the account of the restoration which he based on "An Interesting Account of Several Remarkable Visions and the Late Discovery of Ancient American Records" written by Orson Pratt and published by Ballantyne and Hughes in Edinburgh in 1840. It is my understanding that the English original of Hyde's work no longer exists. There is a re-translation back into English by Justus Ernst of the church translation department, it follows Pratt's wording too closely. Any information about the original must be based on a translation back into English from the German. Of course, this procedure has its limitations but it can nevertheless be of some help.

If in the course of translating Hyde's book back into English, I ran across a scriptural reference, I would open the King James translation of the Bible and use it instead of translating it myself from the German. This procedure insures traditional phraseology and recognition on the part of the reader. It has generally been assumed that this is the same procedure used by Joseph Smith when he came across a passage from Isaiah during his translation of the Book of Mormon. Since my specialty is biblical German, I wanted to check to see which version of the German bible Orson Hyde had used when translating scriptural passages into German. I assumed he would have used the Luther translation, although I do not know whether he knew anything about Joseph Smith's views about the "Old German" translation.

We know two things that can help us in our examination of the text: 1) Orson Hyde had had only limited exposure to German (from "eight days" to a few months at most according to the information we have at present) and 2) a native speaker of German helped him with the translation. Just what the language experience of this person was and how much or how extensive this help was, is not known.
Let's look at two aspects of the language in the book: 1) accuracy, consistency and correctness and 2) translations of passages of scripture within the text.

Technical aspects: 1) spellings no longer in use (eilf = elf, Schoos = Schoß, Verläumdung = Verleumdung, That = Tat, Amerika's = Amerikas, Aéger = Ärger, graviren = gravieren), 2) lack of capitalisation (vereinigte Staaten 13:15, nichts wünschenswerthes 23:22), 3) inconsistencies (-secte/Sekte, dieß/dies, Ereigniße/Finsternisse, grosse/große, müße/müssen, egyptischen/ägyptischen), Týrancë/Tyrannei, Indier/Indianer) 4) misspellings (Jusquehannah = Susquehannah 29:3, erschrack = erschrak 28:12, Nacht 18:10 = Naht, Moroni [41] = Moron, 5) incorrect forms seine Glaube [60] = sein Glaube, Ideen [50] = Ideen sind [76], unter 8 Jahren alt = unter 8 Jahre alt [75], 6) nach = nach Osten, aile Dinge, was = alles, was [59], 7) poor choice of words ('the glory of God' die Herrlichkeit Gottes = die Glorie Gottes [21], 'remnant' Triimmer = Überrest [19:3], 'gross darkness' groBe Finsternis = Dunkel [47], 'approximately' beiläufig (regional) = ungefähr [26:23].

Items 1-4 could be the result of inconsistent typesetting on the part of the printer or of proof reading on the part of the proof reader if there was one, possibly from among the printer's staff or Orson Hyde or the person who helped him with the translation. The others (with the possible exception of beiläufig) are most likely Hyde's responsibility and arise from his close reliance on English.


Here are some instructive examples:

1. dann werden die Vollbringer der Ungerechtigkeit deinen Untergang suchen [25:15]

'... then shall the workers of iniquity seek thy destruction.'

In addition to 'workers of iniquity,' English also has 'evil doer.' Biblical German for the most part has the well established word der Übeltäter, 'the evil-doer.' The phrase Vollbringer der Ungerechtigkeit would likely not be recognized as a biblical phrase.

2. dieser gegründeten Kirche [51]

I believe this phrase is intended to refer to 'of this established church,' as it stands it means 'of this founded, formed church.' It should more properly read dieser staatlich anerkannten Kirche or dieser Staatskirche.

3. Sie [die Platten) wurden hier niedergelegt als Mittel zur Anhaufung irdischen Gewinnes, oder zur Verherrlichung dieser Welt [22:13]

'They were [not] deposited here as a means of accumulating worldly gain or of glorifying this world.'

We know from various other accounts that this is the exact opposite of what is intended. It seems clear that the nicht was omitted and proofreading did not correct it. The presence of oder reinforces this interpretation.

4. Diese beiden letztern Sprachen wurden gleichzeitig von dem Volke unterstanden, als sie von Jerusalem ausgingen [41]

'These two latter languages were both understood by the people at the time they left Jerusalem.'
To me this is a mistranslation. *Unterstehen* regularly means 'to be subordinate to' or 'to be under the jurisdiction of' or in a second meaning 'to dare,' neither of which fits the context. Hyde had fallen victim to what Henry Sweet calls the arithmetical fallacy, that is, he had put together two pieces of German according to the rules of English expecting a correct German result: Eng. *under* = Ger. *unter*, Eng. *stood* = Ger. *standen*, Eng. *understood* = Ger. *unterstanden*. It was close, of course, because English and German are closely related languages but the correct form would have been *verstanden*.

5. Dieser geheiligte Schatz ist von den Gliedern unserer Kirche anerkannt ...

'This sacred treasure is recognized by members of the church ...'

There are numerous passives of this sort with *sein* based on the pattern of English with *be* 'sein' instead of using the standard German auxiliary *werden*.

6. 'ful(fill)'

in der eilften Stunde des Tages zur Vollfüllung alles dessen, was Er seinem Diener Johannes offenbarte auf der Insel Patmos [52:12]

'in the eleventh hour (of the day) for the fulfillment of everything that he revealed to his servant John on the isle of Patmos'

um den Willen seines Vaters zu vollfüllen [66:6]

'in order to fulfill the will of the Father'

die Vollheit des Evangeliums Jesus Christi [22]

'the fullness of the gospel of Jesus Christ'

The word *Vollfüllung* is not found in Campe, Adelung or Grimm. For me all these words are creations on the pattern of English 'fulfill, fulfillment, fullness' and should be replaced by *erfüllen*, *Erfüllung* and *Fülle*.

7. Words for 'priesthood.'

Über das Recht und die Gewalt des heiligen *Priester = Amtes* [44:1]

'Concerning the rights and powers of the holy priesthood'

als wir unter seinen Händen die heilige *Priesterweihe* empfingen [49:13]

'when we received the holy priesthood under his hands'

so will der Herr einigen die *Priesterwürde* verleihen [24:14]

'the Lord will give the holy priesthood to some'

According to Adelung (contemporary dictionary, 1793) *Priesterweihe* is used in the Roman (Catholic) and Greek (Orthodox) Churches to indicate the ceremony consecrating a cleric to become a priest. The usual word in protestant churches is *Ordination*. The other two words also stem from Catholic tradition. None of these words are used in the language of the church today. We use *Priesterium (übertragen)*. See below.

Let us turn next to some of the quoted scriptures and look at some more words dealing with the
priesthood. Here is part of the text of Section 13 in the Doctrine and Covenants as quoted in Hyde’s book.


Compare this version with that of the translation by Heinrich Eyring in 1903 (Dritte Auflage). The first edition by the same translator in 1876 does not contain Section 13:

Auf euch, meine Mitknechte, übertrage ich in dem Namen des Messias, das Priestertum Aarons, [welches die Schlüssel der Erscheinung von Engeln, des Evangeliums der Buße, und der Taufe durch Untertauchung zur Vergebung der Sünden hält; und dieses soll nie mehr von der Erde genommen werden], bis die Söhne Levi's dem Herrn wieder ein Opfer in Gerechtigkeit darbringen.

<table>
<thead>
<tr>
<th>Hyde</th>
<th>1876</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dienstgenossen</td>
<td>Mitknechte</td>
</tr>
<tr>
<td>verleihen</td>
<td>übertrage</td>
</tr>
<tr>
<td>Heiland</td>
<td>Messias</td>
</tr>
<tr>
<td>Priesteramt</td>
<td>Priestertum Aarons</td>
</tr>
<tr>
<td>Gewalt</td>
<td>Schlüssel, Vollmacht, Macht</td>
</tr>
<tr>
<td>damit</td>
<td>bis</td>
</tr>
</tbody>
</table>

Our current Section 13 follows the wording recorded by Joseph Smith and first appeared in Times and Seasons (1 August 1842). Later it was published in the 1876 edition of the Doctrine and Covenants. It was used as the basis for the first German translation in the German edition of 1903. It was not included in the first German edition of 1876. This accounts for the longer version quoted here (in brackets and containing the text "Priesthood of Aaron ..." along with the word until; see HC 1:39).

According to the footnote in the current edition of the Doctrine and Covenants (p. 56, Joseph Smith, History 1) the wording recorded by Oliver Cowdery (shorter with the words this Priesthood and this authority and the word that) appeared in the Messenger and Advocate, vol 1 (October 1834), pp. 14-16. It is obvious from his choice of the conjunction damit that Hyde either had a copy of this version or knew it from memory. Aside from these considerations, Hyde’s choice of words is considerably different from that of the later translation (1903), such that it flavors the entire passage. It may simply be that in the older form it is unfamiliar to us. Dienstgenossen is not a biblical word, verleihen and Priesteramt have a Catholic flavor and Gewalt is too strong (‘force’).

On p. 79-80, Hyde quotes extensive portions of Section 84 of the Doctrine and Covenants. They are enclosed in quotation marks but no section or verse number is given. Verses and portions of verses are omitted, mostly likely for brevity.

2. Jeremiah 17:6 [59]

KJ: For he shall be like the heath in the desert

Hyde: der ... soll gleich der Hitze der Wüste sein
Luther has Heide and Allioli has Heidebaum reflecting the original and English heath. This looks to me very much like a misreading of heat for heath or looking in the wrong place in the dictionary and getting the wrong definition. The word Hitze fits with Wüste but had the translator been following a German translation he would have recognized the mistake.

3. Malachi 4:1 (3:16) [64]

KJ: For, behold, the day cometh, that shall burn as an oven; and all the proud, yea, and all that do wickedly, shall be stubble: and the day that cometh shall burn them up, saith the Lord of hosts, that it shall leave them neither root nor branch.

Hyde: daß der Tag kommen wird, welcher brennt gleich einem Ofen, und daß alle die Stolzen und die, welche Ungerechtigkeiten üben, gleich Stoppeln sein werden. Und der Tag wird kommen, der sie hinweg brennt, Wurzel und Ast, sagt der Herr.

A: Denn siehe, es wird kommen der Tag, entflammt wie ein Ofen, und alle Hoffartigen und alle, die Bosheit üben, werden wie Stoppeln werden; der kommende Tag entzündet sie, spricht der Herr der Heerscharen, und läßt ihnen weder Wurzel noch Zweig.

This verse poses a different problem. I have looked at various translations and have found none that fits very well. I first looked at Luther but it was so different that I didn’t even show it here. Some of the phrases look as though they might have been influenced by English. Most German versions use the words Wurzel and Zweig for ‘root, branch,’ but de Wette (1839) uses the same pair as in Hyde (Wurzel, Ast). The question is whether Hyde had access to the translation of de Wette and borrowed the word Ast, or whether he found the word in his own sources. It is not likely that a native German would come up with a less frequent rendition of this phrase. There are some more hints in this verse that he was translating from his recollection of the English: 1) the omission of the phrase ‘of Hosts,’ 2) hinweg brennt for ‘burn up,’ 3) absence of the negative correlatives ‘neither ... nor.’

4. Matthew 6:9 [82]

KJ: Our Father which art in heaven
Hyde: Vater unser, der Du bist in dem Himmel
A: Vater unser, der du bist im Himmel

In this verse, there cannot be transfer from the King James translation. The word order shows that it comes through the catholic tradition of the Vulgate (pater noster). Hyde can only have access to this through a (probably catholic) native speaker or a catholic translation of the Bible. We remind ourselves at this point that he was staying in the very catholic area of Regensburg.

5. Matthew 10:40 [71]

KJ: He that receiveth you receiveth me, and he that receiveth me receiveth him that sent me.

Hyde: Der, welcher euch aufnimmt, nimmt mich auf, und der mich aufnimmt, nimmt den Vater auf, welcher mich gesandt hat.

A: Wer euch aufnimmt, der nimmt mich auf; und wer mich aufnimmt, der nimmt denjenigen auf, der mich gesandt hat.

So far, I have not found any German translation that uses der, welcher in these very common biblical correlative phrases. The German form is with wer, der; the English (King James) form is he, that. My inference would be that this verse was heavily influenced by the English. There is also an addition which
accords with the sense of the verse but which is not in the original (King James): den Vater. It appears to me that Hyde merged the wording from Matthew with the wording in the Doctrine and Covenants (DC 84:36-37, first published in the 1835 edition), which has the word Father but not the phrase that sent me. I believe this is further evidence that he was either quoting from memory or did not check the wording in the bible.

6. Acts 2:38 [67]

KJ: Repent, and be baptized every one of you in the name of Jesus Christ for the remission of sins, and ye shall receive the gift of the Holy Ghost.

Hyde: Bereuet und lasse ein jeder sich taufen im Namen Jesu Christi zur Nachlassung der Sünden, und ihr sollet empfangen die Gabe des heil. Geistes.

A: Thut Buße und ein jeder lasse sich taufen im Namen Jesu Christi zur Vergebung eurer Sünden; und ihr werdet empfangen die Gabe des heiligen Geistes.

By far the commonest rendition of 'repent' is 'tut Buße,' but there are those that use bereut. I cannot find any that use Nachlassung, which would be possible but unusual. Several use the more common Erlassung. I assume that he was translating from English and not following a German biblical text.

7. James 1:5 [15, 58]

KJ: If any of you lack wisdom, let him ask of God, that giveth to all men liberally, and upbraideth not; and it shall be given him.

Hyde [15]: Wenn Jemand von euch der Weisheit bedarf, so lasst sie ihn von Gott begehren, der da allen (Menschen) freigebig gibt und nichts vorwirft, und es soll ihm gegeben werden.

Hyde [58]: Wenn jemand von euch der Weisheit bedarf, lasst sie ihn von Gott erbitten, welcher allen (Menschen) freigebig gibt, und es nicht vorrückt, und sie wird ihm gegeben werden.

A: Fehlet es aber jemanden aus euch an Weisheit, der erbitte sie von Gott, welcher allen reichlich giebt, und es nicht vorrückt, und sie wird ihm gegeben werden.

I would assume that between the first version and the second version of this scripture, Hyde or his helper had access to a bible translation with vorrückt in it, more specifically that of Joseph Franz von Allioli (1830-37). His translation belongs in the catholic tradition of the Vulgate, Dietenberger, Ulenberg, Allioli and the current Einheitsübersetzung. If you will allow me to go to the next verse, I think we can find verification for this hypothesis.

8. 1 Cor 15:29 [81]

KJ: Else what shall they do which are baptized for the dead, if the dead rise not at all? why are they then baptized for the dead?

Hyde: Was thätet sonst die, welche um der Todten willen sich taufen lassen, wenn es gewiß ist, daß die Todten nicht auferstehen? Warum lassen sie sich für dieselben taufen?

A: Was thätet sonst die, welche um der Toten willen sich taufen lassen, wenn es gewiß ist, daß die Toten nicht auferstehen: Warum lassen sie sich für dieselben taufen?
There can be no other explanation for the word for word correspondence between the texts; our experience with the previous, varying quotations only serves to emphasize the agreement here. There is only one possible point of disagreement (the spelling Todten for more modern Toten and that would probably vanish, if I had an original edition of Allioli (1830-37). There are other verses that show this identity of text: Off 7:2, 14:67, 18:1, Röm 6:4-5. When I discovered this, I immediately jumped to the conclusion that all scriptural quotations coming after page 52 in the text would follow Allioli. This turned out not to be the case. I examined all the quotations again and found that every verse which included chapter and verse as well as quotation marks was direct from Allioli. The others were enclosed in quotation marks but the chapter and verse were not listed. (There were two borderline cases: 1) the quotation from Isaiah [43] (in dem 37ten Kapital seiner Prophezeiung am 31. und 32. Vers) varies the form of citation somewhat, uses the English form of the prophet's name and generally follows the King James wording, 2) the quotation from Thessalonians [90] (Thessal. 4. Kap. 16. und 17. Vers) ends but does not begin with quotation marks, fails to note that the quotation is from the first epistle to the Thessalonians and generally follows the King James wording.) I could now pretty well distinguish between those that followed Allioli and those that did not, but I was still left with the question as to why some scriptural quotations did not follow the Allioli text. One can only guess that access to the text came too late or that the typesetting had already progressed too far, or that there was not enough time to go back and make the changes.

From my examination of the text, I have concluded the following:

1) there are more misspellings, inconsistencies and grammatical errors in this text than in those written by native speakers of German

2) many phrases and word choices are based on English

3) Hyde did not use Luther as a source for quoting scriptural texts

4) for his quotations from the Doctrine and Covenants, he probably relied for at least some portions on printed newspaper accounts or followed a handwritten copy

5) he probably translated from his recollection of the King James wording of the text, or at least he did not check the text closely

6) at some point he or his helper had access to an Allioli translation (1830-37)

7) quotations from the text in Allioli are enclosed in quotation marks as well as chapter and verse (e.g. 1 Cor 15:29)

8) quotations translated from the English (King James) wording are enclosed in quotation marks but do not include chapter and verse

9) the reason or reasons that only some of the quotations are taken from Allioli remain unknown

Lest you think I have been too harsh on Hyde, let me read you his own opinion on the matter. They are found near the end of the appendix. If we cite them in English they of necessity must be a translation from the German [114].

Let no one ever think of holding the principles in this work up to derision or of ridiculing them, for it will not be of any benefit to him or his listeners. I do not claim that this work is technically perfect; I do not understand the German language perfectly, but the principles which it emphasizes are true and good ...
He was also very much aware of the difficulties involved in translating when one is not familiar with the language. In a letter to Joseph Smith dated July 17, 1841 (HC 4:384) he writes:

From past experience I know that the keen edge of any work translated by a stranger, in whose heart the spirit of the matter does not dwell, is lost--the life and animation thereof die away into a cold monotony, and it becomes almost entirely another thing... It appears to me, therefore, that some person of some little experience ought to know this language so as to translate himself, without being dependent on strangers...

Now we see clearly the dilemma Hyde faced: he needed and wanted help with the mechanics of the language but he did not want to leave the work of translating to someone who did not understand the sacred principles of the restoration and who did not possess the 'spirit of the matter.' According to Hyde, the ideal would be someone who had both the linguistic skills and the right spirit. He was willing to accept a few deficiencies in the language, if he could infuse the text with his testimony and an understanding of true principles. If you read the entire book, you will see that Orson Hyde made the right choice.
The challenge of this paper is to say enough about the subject of truth in a short space so that the picture of truth that emerges is not a false witness.

You may be aware that in the long history of the problem of truth there have been some principal answers as to what truth is. The correspondence theory of truth holds that truth is ideas or statements which are perceived empirically to correspond to the nature of the universe. The main problem with the correspondence theory is that empiricism often yields false results. Another historic theory is that truth is the property of propositions which rationally cohere with certain fundamental truths; this coherence would be good if we could only find those fundamental truths. The pragmatic theory of truth says that what works may be taken as true; but what that theory supports is that what works does work, not why it works or what it is that works. A recent entry into the arena is the linguistic theory of truth as initiated by Wittgenstein and articulated by Garth L. Hallett in the book Language and Truth (Yale University Press, New Haven, 1988). This linguistic theory holds that statements are true if they are faithful to the linguistic norms of the culture in which they are uttered. I believe there is a good deal of merit in Hallett's formulation in that he does well represent how the word "true" is actually used in society, but that his theory also falls short by not giving a clear statement as to what truth is and in failing to handle the problem of untruth in ordinary usage.

I therefore now proceed to give my own theory of truth and true, hoping to shed light on this important subject.

I define truth as a synonym for reality. Reality is all that exists, or has existed or yet will come into existence. One cannot discuss reality without making fundamental metaphysical commitments, which I now proceed to stipulate for my ideas of truth.

I understand existence to be composed of material things in various orders, arrangements and functions. These material things and their relationships constitute a whole, each part of which is essential. Thus truth is one, and cannot be divided. To be grasped as truth, it must be grasped as a whole, all that is and was and will be in all of its whys and wherefores, particles, subsystems and totality. Needless to say, this truth is beyond the grasp of any human being.

Each human being is a particular part of the whole of truth, a participant. Each of the feelings, ideas, and representations of a human being are part of the whole truth. The pertinent and pressing question about any given human being is then how he or she represents the truth of the universe to self and to others, and how intelligently one takes ones place in that great truth.

Of principal concern to us is representation of truth. We shall define "true" as a quality of something which measures up to a standard. Thus human beings are true to their word if they do what they have promised to do, and their statements are true if and as those statements measure up favorably to the truth of the universe. What are the possibilities that what an individual thinks or says can be called "true"? To answer that question, a taxonomy of human representations must be posited. We will now explore
a taxonomy which begins with representations which have the greatest possibility of being most true and ends with those least true.

The general label which I give to all human representations of truth is "factitions," from the Latin facere. I use this term to emphasize that in every case, human attempts to characterize truth are for each individual a creative making and doing. Human beings do not passively reflect the universe at any time in their characterizing of it. There is a personal element in each factition which is ineradicable. To use the analogy of a landscape painter, every human factition of truth is an attempt to paint some piece of the universe in a helpful manner. But the painting is never exactly true relative to the truth for at least two constant reasons: first, every human representation is an abstraction from truth, leaving out much that is true; second, no human representation can capture the whole, and only the whole is the truth.

The first level of human representation is perception. Perception, or conocer, kennen knowledge, is the direct sensory inspection of some aspect of the universe. In that direct sensory relationship perception is as close to the truth of things as a human being can get. Sensation is always particulars and of particulars. But this perception is ordinarily flawed by the fact that sensation is not perception until it is interpreted by the mind of the person. That interpretation is done on the basis of the total contents of the mind of the person; all of his previous sensations, ideas, theories, hopes, fears and inhibitions color his interpretation of sensation. Sensations must be read, just as a book must be read, to make any "sense."

The categories of understanding which the person uses to interpret the particular sensations are usually themselves universals. These universals are theories as to what is important and true in the universe and what is not. The more truth the person already has in mind, the more true will be his perceptions. But it is quite safe to say that no human ever perceives all things truly. The best and paradigm case of human perception is found in the direct, continuous, present, proximal sensing of a limited and very familiar aspect of the universe by one who is an expert on that subject. At best direct perception is once removed from the truth, which is to say that the best representation of the truth a human can make may yet be false.

The second degree or echelon of representation is the understanding of an experienced person. This is saber, or wissen knowledge of the world. At its best and surest this understanding is limited to the spatial, temporal, and causal sequences with which the person is very familiar. Identities, differences, continuities, etc., are part of this domain. At its weakest, this type of representation may be so flawed by false theories of the universe as to render the individual without a workable hypothesis as to what is being perceived, as is seen in certain types of mental illness. At best, these representations are twice removed from truth; at worst they are wholly untrue.

The third echelon of human representation of truth is found in the ability to do what one wishes to do. This ability exists only in doing what one wishes to do. This is koennen knowledge, can do in English. This kind of representation of truth comes after perception because the desire to do things comes only after understanding the possibility that they might be done. This can-do knowledge is a representation of truth by emphasizing what works, what the effective sequences of action are that are necessary to produce a certain result. Producing results does give us the truth that a certain action has produced a result, which is a specialized form of understanding, but knowing that a thing has happened does not involve knowing why that thing happened. Thus a full understanding of echelon two is a better representation of truth than the partial understanding of what works as found in echelon three. And echelon three is thrice removed from the truth.

Perception, understanding and the ability to do something are personal representations of truth within the individual. They have been the inspiration for the correspondence,
the coherence, and the pragmatic theories of truth. Though not truth, they are the representations of truth closest to the truth and therefore the most true ideas which the individual may have. They are not linguistic, but they reflect heavily the prior linguistic experience of the individual. The remaining categories of representation of truth by persons are all linguistic functions.

The fourth echelon of human representation of the truth is found in the individual's witness of his own perceptions. Using his own personal perceptions as a base, the person formulates some verbal means of expressing a new perception. All words represent universals. When an individual tries to express the particulars of his experience in words he always faces a mismatch between what sensations are and what words can do. That problem, compounded with the universals of interpretation and understanding which color all perception, make an individual's testimony as to what he has personally perceived four times removed from the truth.

The fifth echelon of human representation is in the witness an individual gives of his understanding of actual experiences he has had. All of the problems of perception and the reporting of perception are here augmented by the potential flaws in his understanding. A person might honestly report a temporal or spatial or causal sequence which he has observed, but be so thoroughly mistaken as to what actually was happening as to be a totally misleading witness. This fifth echelon is five times removed from the truth.

The sixth echelon of human representation of the truth is in the individual's linguistic representation of what has worked for him as he has tried to fulfill his objectives as a person. Colored by his perceptions and limited by his understanding of the truth, this echelon is further hampered by the fact that when an individual is successful in accomplishing something he seldom can give an exhaustive account of all that he did and of all that the environment furnished to bring about his desired result. The individual knows that in situation X he did Y and obtained Z, but cannot give a full and accurate account of X or Y or Z. Therefore, this sixth echelon of representation is six times removed from the truth.

The seventh echelon of human representation is human witness as to inductive generalizations he has made about the world out of his own experience. We have now crossed the line from the possibility of inadvertent error in representing truth to the overt and deliberate embellishment of what the individual has experienced. In other words, we are now in the realm where pure guesswork characterizes the attempts of the individual to represent the truth. All interpolations and extrapolations are technically guesses, and these guesses suffer even more from the possibility of wishful thinking than do the previous levels of factitious. Valuable and useful as some inductive generalizations of experience may be, such representations are at least seven steps removed from the truth.

The eighth echelon of representation is theory. Theories are understandings that are deliberately invented to characterize some aspect of truth which cannot be the subject of direct empirical observation. Thus discussion of the nature of atoms, of space-time matrices, of how man came to be on the earth, of what is good and evil, --all such are inventions of men to try to overcome their lack of ability to see for themselves the truths of these matters. All historical accounts and all interpretations of linguistic formulations are types of theories. This echelon includes all quotation of other human beings. While it is true that logical consequences of a theory sometimes offer the possibility of empirical confirmation, no empirical experience necessitates either the adoption or the rejection of any theory. Theories are often accepted and rejected on non-experiential criteria. Theories are eight times removed from the truth.

The ninth echelon of human representation of truth is found in overt fictions. These are counted as representations of truth because one main use and value of fiction is to
present ideas as to the way things really are in some respect using non-historical characterizations. These characterizations are usually attempts to present inductive generalizations or theories of truth in an artistic form, one that is pleasing or attention-getting. But as representations of truth, fictions are at least nine steps removed from the truth of things.

The tenth and final echelon of human representation of the truth in this taxonomy is found in the deliberate lie. This lie is a deliberate mis-representation which is known to the positor of the lie to be a lie but which he hopes he can get other humans to accept as true, as adequately representing truth. Lies are very effective in a world where truth is important and valued, where truth is difficult to come by, and where people are not always very careful as to what they accept as a representation of truth. Such is the world in which we live. Thus lies are ten steps removed from the truth. But they are not very far removed from those representations which are close to it in the echelons of representation.

Sometimes human beings do recognize the importance of truth and take special precautions to try to eliminate falsehood from linguistic exchanges. In law there is a recognition that the personal testimony of an eyewitness to an event is more valuable in establishing the true representation of an historic event than any other kind of representation, and that the testimony of several witnesses is better than that of only one. Also recognized is the testimony of expert witnesses, who are allowed to tell of their understanding and can-do knowledge, sometimes even of their inductive generalizations and theories. But since that kind of representation is from four to eight times removed from the real truth, the justice of our courts of law sometimes miscarries because it must accept such a poor representation of the truth as this, for want of better. The scholarly world recognizes that primary sources (fourth echelon representations) are much better evidence of the truth than are secondary sources (eighth echelon representations).

Science as an institution has sought to rid itself of the problem of representing truth by eliminating all personal knowledge and witness of truth, the first four echelons, and by replacing them with inductive generalizations and theories which are agreed upon by the majority of competent scientists. Science thus focuses on the seventh and eight echelons of truth representation. Scientists essentially say to the rest of mankind: We will manage your truth concerns for you; just put your trust in us and we will deliver you from error, because anything different from or outside of what we propound is error. Historical insight reveals that science is not omniscient but advances by replacing one scientific representation by another through time. The power of science is of course not in its representations. Its power and prestige come ultimately from the fact that the technology associated with modern science is formidable. Science is accepted as a painter of truth because of the fireworks it can produce. Producing fireworks does show that sometimes the inductive generalizations and theories of science do have some positive relationship to the truth.

Art in some of its forms is a non-literal attempt to represent truth, as discussed above in the matter of deliberate and overt fictions. Another side of art is that it attempts to capture ideals in artistic production is the attempt to "realize" things which are taken to be true, good and beautiful. The question about such art is, does it fully embody the ideal which the artist set out to create? Inasmuch as an artist does create, his artistic production becomes truth, part of the whole being of truth, which itself must and may then be represented by some one of the above delineated ten echelons of human representations of truth.

We come now to some conclusions and applications.
1. Truth is a whole and cannot be represented adequately by human beings. Therefore a large measure of humility is appropriate in every human attempt to find or state something which could be called true.

2. There are no degrees of truth. Something is either the truth or it is not. But human representations of truth certainly do come in degrees, in at least the ten steps of removal from the truth as explicated in this paper. The trueness of a representation is thus a qualitative variable which may vary from 1 to 10, 1 being best. But human beings have no human means of being sure that their representation of the truth is true. Error always lurks as a real possibility.

3. There is also a quantitative measure of truth as well as a qualitative measure. How much truth a human being represents is a function of the amount of experience he has had with whatever fraction of the universe he has experienced.

4. All human representations of the truth are creative, factitious, and are therefore as much a measure of the artificer as they are of the truth being represented.

5. It is easier to know truth, to represent it to oneself, than it is to speak truth, to represent it to others.

6. Most of human discourse, statistically speaking, lies at the untruth end of the spectrum rather than at the truth end.

Which brings us to the necessity of including in what we say some mention of spiritual matters. Spiritual matters are part of the reality of the universe, and to try to discuss truth without saying something about spiritual experience would be deliberately to falsify everything that has been said. There are two troublesome problems that must be dealt with in connection with spiritual matters. One problem is that every human being is more an expert on his own spiritual experience than is any other human being. This is good in that it fosters individual initiative and independent thinking. The other problem is that because there are two spiritual sources, many persons latch onto a spirit that fosters untruth, and in their independence, are difficult to assist. A typical human attempt to overcome these problems is to encourage people to denigrate all spiritual experience in favor of trusting in some human authority. We shall show that that is a poor expedient, if getting close to the truth is the goal.

The individual in his own personal experience of truth can be closer to the truth than any linguistic and socially acceptable account of the universe could ever be. Personal experience is always spiritual, and furthermore each honest person knows that there are at least two spirits besides his own which affect him constantly. Let us then make a brief account of truth in light of those two spirits which affect human beings.

One spirit is the spirit of truth and the other spirit is a lying spirit. By whatever names these spirits are known to men, they are known to men. Whenever a person attempts to characterize the truth, to know it or to speak about it, one or both of those spirits is at hand to assist in the process.

It is the mission of the spirit of truth to assist the person to see, to understand, and to be able to do all that he needs to do in this world. But the spirit of truth is not primarily interested in truth. What the spirit of truth is more concerned about is righteousness, doing good in the world. Truth is a means to doing good, but knowing truth is never more important than doing good. So the spirit of truth comes to a person first to tell them the importance of doing good, then to tell them what truly is the good to be done by them in their situation, then to tell them any other truth they need to know to be able to do the good they should do. Should what that person needs to do to do good involve linguistic characterizing of the truth about the universe for the benefit of another human being, the spirit of truth will instruct the speaker as to what to say,
and then will interpret for the hearer, so that the exact portion and quality of truth necessary for both the speaker and the hearer to do good will be communicated.

The lying spirit is of course also not principally interested in truth and error. That spirit is principally interested in getting human beings to do evil to one another, to damn and hurt one another. The chief weapon of this spirit is lies, thus this is a lying spirit. He will tell truth and will influence human beings to know and speak truth whenever that will bring about evil, and he promotes lying whenever it will bring about evil.

So if a human being understands the difficulties of representing truth and also knows these two spirits, how can or should he or she act? We shall first delineate the case of the follower of the spirit of truth, and then the case of the person who follows the lying spirit.

How will a follower of the spirit of truth act in this world? Such a person will seek to feel the influence of the spirit of truth in all situations. He or she will be apt to listen to and quick to do that good which that spirit of truth commends, seeking also to gain true perceptions, true understanding, and true ability to do that which needs to be done. Should this person need to speak of the truth, he or she will assiduously strive to measure every gesture, word and characterization to itself become a good and a true representation, acting and speaking as humbly as possible under the influence of the spirit of truth. When one speaks by the spirit of truth, though words cannot convey the truth, the truth of the matter can be manifest to the hearer by that same spirit of truth by which the speaker speaks. Thus it is the spirit of truth that is responsible for the truth, not the human speaker. This does not give license for the speaker to be careless with the truth, for he must attempt always to speak truly, by the spirit of truth. But truth is yet the province of the spirit of truth.

Should the follower of the spirit of truth encounter the words of another human being who speaks by the spirit of truth, that hearer will pay close attention to the personal witness of particulars which the speaker relates out of his own experience. If the matter is important, the hearer will go to see for himself. He does not want to depend on the word of another, even a good word, because words are always further removed from the truth than is personal observation under the influence of the spirit of truth. Should the good speaker speak of things not in his personal knowledge, that person will speak only under the influence of the spirit of truth, and the hearer will then apply to the spirit of truth to receive a personal manifestation of the matter from the spirit of truth for himself. He knows that personal knowledge is always closer to the truth than a manifestation reported by another, even if the speaker is truly saying what he has been led to say by the spirit of truth. Thus the influence of the spirit of truth is to cause every person to seek to know for himself both the natural things he may observe and the unseeable things concerning which he may receive his own personal instruction from the spirit of truth.

When one who hears by the spirit of truth hears a person who speaks by the lying spirit, the results are much the same. The hearer will not accept the reported personal knowledge of the speaker, but will go see for himself. Neither will he accept the witness of things which are not personal knowledge, but will seek further from the spirit of truth the truth about the matters on which the person of the lying spirit speaks.

What happens when one of a lying spirit hears another who speaks by the spirit of truth? In this case the person of the lying spirit will accept whatever is in the personal knowledge witness that the speaker gives which the hearer finds to be useful or pleasing, and will reject the rest. The person of the lying spirit hears the speaker who speaks of unseeable matters by the spirit of truth in such a way as to reject what is said unless it can be twisted or interpreted to become pleasing or useful to the hearer.
When one of a lying spirit hears one who speaks by a lying spirit, the witness of personal knowledge is again accepted if it is pleasing or useful. But if the hearer wants to use that knowledge to accomplish something in the real world, he will go find out the truth of the matter by his own personal observation, for even liars must abide truth in that which they wish to accomplish. But in the matters which are not the personal knowledge of the speaker, the hearer of the lying spirit will hear what pleases himself or what he will find useful in promoting lies with others.

Now for some conclusions and generalizations about spiritual matters related to truth.

1. A person of the spirit of truth wants the real truth no matter how unpleasing it is, because only the truth enables him to work in a real way to solve the real problems with which he is confronted.

2. A person of a lying spirit must leave that lying spirit and seek truth to be able to do anything in the natural world, for nature cannot be flattered into cooperation by lies as people can.

3. People who speak truly by the spirit of truth will often be rejected by those who hear with the lying spirit, because the truth does not please them. If truth pleased them, they would seek and hold to the spirit of truth rather than the lying spirit.

4. Persons who seek influence in society by the lying spirit only need to tell those who hear by a lying spirit what pleases them in order to gain power.

5. No person can assure any other person of the truth. That is the domain of the spirit of truth.

The conclusion of the matter is then that two factors must be accounted for by one who would make truth his standard. First he must be more interested in righteousness than he is in truth, for then he will be able to find the spirit of truth and to hold to abide in it without error. Second, he must understand the difficulties and problems in knowing and speaking truth, so that he will believe and speak only by the spirit of truth, and not be tempted to let go of the spirit of truth and propound on his own as if he were some sort of non-human paragon of truth. For to propound on our own that which pleases us is to have fallen into the arms of the lying spirit.
METATEXT: COMPUTERIZED MATERIALS FOR THE STUDY OF SHAKESPEARE'S LANGUAGE
Mark L. Reynolds and Camille S. Williams
Brigham Young University

INTRODUCTION

During the past decade Arthur H. King has compiled detailed sociostylistic materials focusing on the social and dramatic function of language in Shakespeare's plays. His approach views the Shakespeare text itself as a source of historical data; a careful application of sociolinguistic methods can make those data available for a full range of interpretive approaches.

In addition to his own work, he has drawn on the work of others studying Shakespeare's vocabulary, pronunciation, syntax, and metrics. This compilation, if easily accessible to researchers, teachers, actors and critics could significantly aid them in their work.

Since August 1988 we have been working on the prototype of MetaText, a new form of hypertext designed to allow the user to view multiple types of parallel data displayed automatically with the movement of the cursor through the Shakespearean text. Initial testing suggests that computer access to this large database is far superior to accessing it via the printed page.

This paper discusses Dr. King's approach to the study of Shakespeare's language, the categorization and preparation of the data and the various functions and capabilities of the MetaText software.

We welcome suggestions about the content and formatting of the material and about applications of the MetaText to other major texts in literature, philosophy, religion or other fields of study.

RESEARCH ON THE LANGUAGE OF SHAKESPEARE

Though there are a few standard reference books on the language of Shakespeare (such as Abbott's A Shakespearian Grammar, and Hussey's The Literary Language of Shakespeare), there is no encyclopedic work with language as its focus. Numerous studies of Shakespeare's language are published in dozens of periodicals, some serving the interests of linguists, others serving as a basis for critical interpretations. It is a huge task simply to sort through
all the articles containing references to Shakespeare's language. Many studies are restricted to one short passage, one aspect of style, or to one poem or play. Though these restrictions are defensible, this means that a few small segments of text have received intense scrutiny and very large segments have received little attention in the published body of work.

Stylistics is in a state of flux; stylisticians use methods and theories ranging from transformational grammar to traditional rhetoric to semiotics. Because stylistics or linguistic stylistics, or linguistic criticism are terms used to name diverse activities in which linguistic skill is used for textual analysis, it is necessary to state the premises underlying the stylistic approach to Shakespeare we've used.

HISTORICAL SOCIOSTYLISTICS

Dr. King's approach analyzes Shakespeare's use of language within the contextual norms of Shakespeare's time. This requires reference to historical linguistics, sociolinguistics, and contemporary linguistics. This method may be termed historical sociostylistics. The emphasis is on the social and dramatic function of the language in the specific context of a play.

Some stylistic work attempts to isolate a subset of language with definable characteristics that make it 'poetical,' or 'literary,' rather than 'ordinary' language. Many studies attempt to isolate the characteristics of an individual writer's 'style'.

This approach attempts neither. Instead it is expected that literary language will reflect the conventions of the society in which it was produced, and in the case of Shakespeare's drama, will reflect the language use of his day as mediated by contemporary social and literary conventions. It is also posited that there exists no set of language characteristics comprising a 'Shakespearean' style; rather, it is expected that Shakespeare used a variety of registers and styles which helped his audiences make social, and hence dramatic, sense of the roles on the stage.

Dr. King's approach analyzes virtually every word of a play in the context of usage across all of Shakespeare's works. In this way, his conclusions are based on the examination of a much larger corpus of material than the short poems, or short prose selections used for most stylistic study. His approach also considers more aspects of language than do most studies, which are limited, for example, to syntax, or to subsets of Shakespeare's vocabulary, or to the use of one rhetorical device. He looks for the cumulative effects of several aspects of language.
Dr. King uses the language data to assess sociostylistic norms, their ranges, and deviations from them. Such norms are not related to the development of a standard English, nor are they based primarily on the statistics of usage, but are inferred from usage in the social contexts represented in the plays.

To summarize, we've spent years gathering information relevant to understanding the social or literary norms for language use and have then used that information to explore, appreciate, and interpret the text.

MATERIALS FOR THE STUDY OF THE LANGUAGE OF SHAKESPEARE'S PLAYS

Because sociostylistic information can help us better understand the plain sense, register, organization, and tone of specific passages, this approach can be helpful in research and in teaching. Since 1975 four of Shakespeare's plays (King Lear, Hamlet, Othello, Macbeth) have been studied intensively; preliminary work has begun on an additional six (The Tempest, Coriolanus, Antony and Cleopatra, Twelfth Night, Measure for Measure, A Midsummer Night's Dream). This study has produced thousands of pages of data about Shakespeare's language in these plays. These data could be "mined" for a variety of instructional or research purposes.

Until now, the information compiled for this approach has consisted of books of notes about the language. Each play studied has been divided into passages 15-20 lines long; in each passage the lexis, syntax, patterns of repetition (rhetorical devices, sounds and meter) have been examined, and their cumulative effect within the passage assessed. The passage division generally corresponds to relevant language characteristics in the text; for example, the 20 lines may be a set speech by one role, or an interchange between two roles; frequently there is a shift of register or tone marking the break between passages. Usually each 20-line passage yields ten to fifteen pages of stylistic information.

PROTOTYPE DATABASE

For the MetaText prototype, we chose to work with Lear 1.1.1-120. Dr. King dictated new material, and we worked to derive a format from the new material that would also allow us to incorporate in the final MetaText program material generated previously. The challenge has been not only to produce and edit the material, but also to collect ancillary material supporting the analysis (or contradicting it), to standardize the format of the material, and to adjust both the program and the format to improve user access.

The MetaText prototype allows the user to choose the types of language information to be displayed, while screening out
the rest. In order for the user to decide what he does or does not want to see, we had to categorize the material. We have used the following codes and categories in the MetaText prototype:

<table>
<thead>
<tr>
<th>CODE</th>
<th>CATEGORY</th>
<th>CONTENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO</td>
<td>Commentary</td>
<td>Interpretative comments on specific lines</td>
</tr>
<tr>
<td>CR</td>
<td>Cross Reference</td>
<td>Material from other works, other authors</td>
</tr>
<tr>
<td>DP</td>
<td>Dramatis Personae</td>
<td>Commentary on names of characters</td>
</tr>
<tr>
<td>IP</td>
<td>Interpretation</td>
<td>Mini-essays, interpretations based on language detail</td>
</tr>
<tr>
<td>LE</td>
<td>LExis</td>
<td>Information about words, phrases, usage; identification of lexical figures such as sense play, polyptoton</td>
</tr>
<tr>
<td>ME</td>
<td>MEter</td>
<td>Four-level scale: unstressed, light stress, normal stress, emphasis = word, phrase, sentence level stress mapped over traditional meter</td>
</tr>
<tr>
<td>RF</td>
<td>Rhetorical Figures</td>
<td>Selected traditional terms (anaphora, epistrophe, etc.) used; no trope/scheme division</td>
</tr>
<tr>
<td>SD</td>
<td>Stage Directions</td>
<td>Discussion of staging</td>
</tr>
<tr>
<td>SO</td>
<td>SOUND</td>
<td>Phonemic/phonetic symbols unavailable for this version</td>
</tr>
<tr>
<td>SY</td>
<td>SYntax</td>
<td>Construction, register, figures (triads, antithesis, etc.) identified</td>
</tr>
</tbody>
</table>
These data can help the user evaluate register, style, conventions and tone; it may make some of his work in the OED and other reference books more efficient, or perhaps unnecessary. Entries are in standard format, and the final version of the program will allow access to any portion of the entry.3

METATEXT SOFTWARE

Dr. King's language material has been used in classes and circulated at conferences for years; it has been criticized as too hard to understand. One reviewer said, "This looks like a book of footnotes; why don't you append some essays?" We've taken some of his advice. Since research questions motivate scholars to use reference works, we wanted a program which would show material from the categories relevant to the question, but "hide" the rest. We also wanted Dr. King to explain at various points his interpretation of the language detail.

Further, the program should be simple to use. Hypertext, in its most common format, frequently asks the user what it should do and which material it should display. Being asked a question while pursuing a train of thought is distracting, so we were dubious about using standard hypertext for this project.

We think that in many inquiries, more than one kind of language information will be immediately relevant (for example syntax and lexis could both be required in studying changes in usage), so the program must be able to display multiple categories simultaneously.

Word searching in the Shakespeare text, as well as in the language material itself, must be a standard option to allow the user more scope in designing his own research strategy. Additional reference works should be included, and should also be searchable. The scholar should be able to make notes and extract textual material without leaving the program, so a word processing and printing function must be available.

With these requirements in mind, we surveyed existing software and found that nothing currently offered would adequately present Dr. King's work for electronic publication. Accordingly, the design for MetaText was drawn up and the prototype was constructed. In addition to the demonstration of the software at the DLLS conference in March 1989, the program and a sample selection of commentary
on the first 120 lines of Lear were shown at the annual meeting of the Shakespeare Association of America in Austin, Texas in April 1989. Several scholars from other fields have also reviewed the prototype.

Their response has been generally favorable; several reviewers' suggestions have been incorporated to improve the human interface, making the program simpler to use. We've also discussed with reviewers which texts or subjects might be effectively published or presented for classwork via the MetaText software. Obviously the software has potential apart from the Shakespeare database.

METATEXT SCREEN AND FUNCTIONS

A copy of the MetaText screen is shown below in figure 1.

![MetaText Screen](image)

Fig. 1. The MetaText screen focuses attention on the Shakespeare text, while allowing the user to see the types of data available for display.

Name, Date, Time. The box across the top of the screen contains the MetaText name, the date and time of use. In future versions, the name of the play, act, scene and line number will also be listed.
Status Bar Display. The single line reverse video window which appears on the bottom line of the Name, Date, Time window displays the abreviations of all categories of comment which are available at any character position. Comments may be assigned to any character, word, line or group of lines. At any time, the user can access the available language data by pressing the view key, then typing in the two-letter category code for a full screen window presentation of data in that category linked to the position of the cursor in the Shakespeare text.

Inquiry Text. The larger box beneath contains the text of Shakespeare's King Lear. The line numbering on the left is that of the Riverside edition of Shakespeare.

Cursor Bar. The bar across the middle of the text window is a reverse video presentation of the line of text selected by the user. The underline character shows the position of the cursor within the line. Pressing a cursor movement key moves the text behind the bar up or down rather than moving the bar itself. This means that the selected text will always appear in the middle of the screen, allowing the user to define fixed windows elsewhere so data can be shown without hiding the the portion of text the data address.

Automatic Display of Data. The user chooses the type of language material to be automatically displayed, designs the size of the window, and positions it on the screen. When the user places the cursor at a point in the Inquiry text for which that category of material is available, relevant material is automatically displayed without further commands. Figure 2 (below) illustrates the automatic display of syntactic data.

The smaller text window at the upper right, overlaying the Shakespeare text, displays a portion of the syntactic data linked to the cursor position. The size of the window limits the amount of material that can be seen initially. For some entries, that space will be sufficient. When it is not, the user can press the view key, type in the the two-letter code to see the material displayed on the full screen.

Full cursor motion is possible within the large screen to facilitate the viewing of the data. When finished with the data in the full screen window, the <escape> key returns the user to the Shakespeare text.

Selecting and Positioning Windows for Automatic Display. Data windows will be automatically displayed only when selected by the user. By pressing the select key, she chooses from the list of categories those she wants displayed automatically. A box of default size and shape is
then presented; she can change the shape and position of the box(es) to suit her needs. Pressing the <enter> key returns her to the list of categories. Any combination or all of the categories may be selected for simultaneous automatic display.

After leaving the category selection menu via the <escape> key, the user sees automatically displayed in the windows data from the categories selected as she moves the cursor through the text.

At any time the user may adjust the shape or position of any window by pressing the display key and making the desired changes.

Word Processing. By pressing the w key, the user may enter the built-in word processor to make notes. The notes may then be filed in separate files, linked to text and automatically displayed in the user's comment window.

Searching. The final version will include full word searching capabilities allowing searches of the Shakespeare text, Dr. King's data, additional reference material and the
user's notes. Statistical information will be available with the searches.

Audiovisual Capacities. When the software is completed, audio and audiovisual material may also be linked to the text allowing readings or productions of Shakespeare to be accessed in the same way other comments are accessed in the prototype.

Updating the Database. Finally, the finished MetaText will allow comments to be made on comments for any number of levels. For example, if a scholar disagreed with Dr. King's assessment of more affected . . . than as a set phrase, we could incorporate that critique by linking it as a comment available from the full screen display of the syntax entry. We feel that access to Dr. King's material will encourage further language work by other scholars, and that some may want to contribute to this database as a form of electronic publishing.

Size of the Database. The only parameters restricting the size of the MetaText will be the storage and memory of the computer on which it is run. For Dr. King's Shakespeare MetaText, it is expected that a laser disk will be required to allow his approach, including relevant published articles and monographs as reference material, to be presented appropriately in an encyclopedic fashion. In an effort to keep the size of the database manageable, and because of the amount of work done on this author, we have restricted our work to Shakespeare's language.

In summary, the MetaText program allows a researcher to selectively view reference material about a given text. Once the windows for automatic display have been selected, shaped and positioned, the user's only input will be to move the cursor through the text, asking for full screen presentation of data when necessary. Such simplicity will allow him to think about his subject rather than the functioning of the program. We think the Shakespeare database presented via the MetaText program will provide a significant reference tool that could aid others in their research for years to come.
APPENDIX: STANDARD FORMAT OF DATA

CO (COMMENTARY)

Arthur H. King's commentary on the text re: pragmatics; tone, style, implication, juxtaposition, omission, etc.

1. Scope or item comment concerns [if it concerns a piece of text, the format is text from the play (line #)]
2. Comment

Terms likely to be used include:

Arcadianism
brevity
colloquial
complement
copy
dramatic irony
Euphuism
register
Senecan

CR (CROSS REFERENCE)

Long references are outlined so you can find the portion that interests you without scrolling through pages of text.

DP (DRAMATIC PERSONAE)

1. Character
2. Comment

IP (INTERPRETATION)

Interpretations are accessible via an outline. The form is:

Title
Mini-essay
Notes
LE (LEXIS)

Format: (some sections will be blank)

1. Word or words from the play (line #)
2. Label (e.g., neologism, normal register, etc.)
3. Definition (includes information from the OED, other reference books)
4. Citations format: <Speaker> [as in WordCruncher], quoted passage, (reference--OTH 3.4.12-15 [play name in all caps]), with each reference separated by semicolons)
5. Basis for conclusion, statistics for usage, etc.

TERMS

abstract for person    pathetic fallacy
affected              personification
archaism              poetic diction
dialect               polyptoton
epithet us ornans     pun
metonymy              sense play
neologism             soria smus
nonce word            word chain
normal                word link

ME (METER) The four point scale: normal(/), light stress (\); unstressed (u) emphasis (/\).

Format:

1. Line, Segment scanned.
2. Label (e.g., isocolon, trisyllabic pronunciation, etc.)
3. Explanation/Function/Examples

TERMS

contrast stress
couplet
disyllabic pronunciation
elision
emphatic stress
hypermetrical unstressed syllable
inverted first foot
isocolon
light stress
phrase-level stress
sentence-level stress
syllabic variation of /l/r/n/m/
trisyllabic pronunciation
word-level stress
RF (RHETORICAL FIGURES)

Format:
1. Words (lines #–#)
2. Label
3. Function.

TERMS

anadiplosis
epanalepsis
anaphora
epanalepsis
anaphora
epistrophe
antimetabole
epizeuxis
chiasmus
stichomythia
climax
symplece
epanadiplosis
tmesis

SD (STAGE DIRECTIONS)

1. Text of stage direction
2. Comment

SO (SOUNDS)

Phonetic/phonemic symbols are not available for this version.

Format:
1. Words (lines #–#)
2. Label
3. Function
4. Similar examples
5. Comment/Effect

TERMS

alliteration, sequential alliteration, vocalic alliteration
antimetabolic sequences
assonance, sequential pairs, subassonance
binders, word-, phrase-, line-, passage-, dialogue-consonance, sub-, super-, consonantal clusters
homoeoteleuton
line patterns
paromoeon
rime
sonance
SY (SYNTAX)

Format:

1. Range: text here . . . to here (lines #--#) or this text (line #)
2. Label
3. Paraphrase/sense
4. Citations: <speaker> quote,(reference) [same as LE]
5. Explanation

TERMS

absolute construction
accumulation
anacoluthon
antithesis (antithetical)
aposiopesis
asyndeton
double negative
elliptical construction
epanorthosis
episodic sentence
ethic dative
graphic present
historic present
hyperbaton

interrogation
inversion
oxymoron
parallelism
parenthesis
parison
periodic sentence
periphrasis/periphrastic
periphrastic do
prosiopesis
schesis onomaton
sententia
stichomythia
transferred modifier

VR (VARIANT READINGS)

1. F reading
2. Q reading
3. Comment
NOTES


2. The amount of stylistic information produced varies according to the text itself. Commonly occurring features, such as variation in the use of thou and you, or use of -eth verb forms are not routinely listed, but are noted when their use is stylistically significant in the context. Language information that is generalizable, such as statements about the use of periphrastic do are made in a glossary of terms to be available by hot key in the MetaText program.


A user interested in only one constituent of the entry could choose to have only that portion displayed, making it easier to fit into the windows data relevant to the user's task, while screening out the rest of the entry.

4. Existing hypertext programs were reviewed, but proved inadequate to the needs of our project. Traditional hypertext is a method of linking a large variety of data types to each other for recall. Since any data object may be linked to any other, the database may be built up using the intuitive associations common to human thought rather than the rigid structures employed by computer scientists in business and record keeping databases.

To use a hypertext program, the user initially finds himself looking at a screen of text. Buried in the text are icons of some sort which are not textual symbols. These represent a link to other data in the database. By moving the cursor to the icon, the user obtains access to a menu of choices. Frequently there are links to more than one piece of data. By selecting the desired option from the menu, the user may view the data which has been associated with the primary text icon.
REFERENCES


THE REACCENTUATION OF FRENCH TEXT

Deryle W. Lonsdale
ALPNET

OVERVIEW

Accented letters serve an important role in French orthography. This paper begins with a brief description of the function of accents in French.

Although the learner of French, native or otherwise, usually learns the correct use of accents through many dictation drills, the accentuation of French text is not always a clear-cut issue in practice. Samples of different French text types reveal inconsistency in the accentuation of upper-case letters, disagreement on the alphabetical order of accented letters, and varying degrees of disregard for accents. An examination of these and other uncertainties of French accent usage will be made.

While the absence of accents in French text does not usually inhibit the human's understanding, there are cases when accents can play a major role in a wide range of natural language processing applications. The motivation for providing correct accent information to such applications is presented.

Finally, this paper investigates the development of on-line "reaccenting" methods capable of re-inserting accents into French text where they may be absent. Different methods for accomplishing this task are presented, and conclusions about their relative computational and practical merits are drawn.

THE FRENCH ALPHABET AND ACCENTS

The French language uses the Roman alphabet with the same 26 letters as English has, as well as the following 13 accented letters:

"â", "è", "ù" with the grave accent:
"é" with the acute accent:
all vowels with the circumflex: "â", "ê", "î", "ô", "û"
"ç" with the cedilla:
"ê", "î", "ù" with the dieresis: "ê", "î", "û".

These letters constitute the standard 39-letter alphabet accepted by most modern French lexicographers.
Some other accented letters do occur in French, on words such as "cañon", "sertão", and "Ångström", borrowed from other languages. We will not consider these exceptional accents; nor will we regard the dots over "i" and "j" as accents, though some experts do.

FUNCTION OF ACCENTS

French accents have been classified into three functional types: phonographic, morphographic, and logographic. A phonographic accent phonetically alters its associated letter. A morphographic accent is used to distinguish similar morphological inflections of the same base form. Logographic accents, too, serve to differentiate otherwise indistinguishable words, but in this case unrelated ones.

The cedilla plays a purely phonographic role, as do the dieresis and the acute accent. They all have an effect on the pronunciation of the letter. An interesting function of the acute accent occurs in some subject-verb inversions; when "je pense" is inverted, for example, the structure is written "pensé-je".

The circumflex has all three roles:

1) phonographic, lengthening the vowel sound, as in:
   "pâte, frêle, maître, ôter, voûte"

2) morphographic, especially to distinguish between subjunctive and indicative verb forms:
   "fut, fût"
   "eût, eût"
   "tint, tint"
   "dites, dîtes"
   "prit, prît"

   but also possessive adjectives and pronouns:
   "votre, vôtre"
   "notre, nôtre"

3) logographic as in:
   "du, dù"
   "forêt, forêt"
   "matin, mâtin".

The grave accent is, for the most part, logographic, playing an important role in the discrimination of high-frequency words which would otherwise be homographs, such as:
When over "e", the grave accent usually induces a more open articulation, and is therefore phonographic in cases such as:

"tréfle, dès, père"

ACCENT USAGE

French orthography is difficult, and the Académie Française has been invaluable in defining proper orthographic use, accents included. On the other hand, there are still several problem areas in French accent use which are seldom, if ever, addressed in the literature.

For example, the Académie has not published its authoritative dictionary since the eighth edition of 1932-35. Only a few lexicographers or dictionary publishers have dared further the official position by proposing and implementing their own up-to-date set of guidelines for accent usage.

The correct accentuation of borrowed words from other languages has always been problematic. For example, the Académie accepts "à priori", but almost nobody uses the accent. "Speculum" can often be seen with an acute accent over the "e", but not always. While foreign words continue to enter into French, foreign word accentuation patterns remain ambiguous, and writers often follow their own preferences and inclinations.

The alphabetization of accented French is another problem. The dictionary industry has apparently not yet reached a consensus on the alphabetic order of accented letters. For example, Robert² and Littré³ list these words in order: "côté, côte, coté, côté" while Larousse⁴ orders them "côté, coté, côté, côte". Similarly, Robert and Littré show "gène" followed by "gène" while Larousse orders them in the opposite fashion. There seems to be no published account of these decisions, much less a standard in general for French alphabetization.

While several French language reference books discuss accent use in general terms, very few endorse or proscribe the use of accents with uppercase letters. In France and Francophone Europe in general, the tendency is not to accent capital letters. Nevertheless, there is little documentary support for such an approach. In only one French orthography manual⁵
have I found such guidance. It mentions that the cedilla must be used on the uppercase "C" even when it is handwritten, but that otherwise one should avoid using accented capital letters, except perhaps when the results could be confusing.

On the other hand, the French Canadians tend to accent their capital letters more often, again with little documentary justification. Jean Darbelnet states⁶ that "Â" is used by some printers, the circumflex is often left off "O", and "È" is justified in all cases. In another reference work⁷, a Canadian typesetter states that one must use all accents on capital letters in French.

There are abundant examples illustrating the inconsistent use of French accents on capital letters. One need only consult daily, weekly, or monthly newspapers, newsletters, and magazines.

Another related, and curious, phenomenon occurs in the formation of French acronyms. Even in instances where an accent is used on a capital letter in an expanded name, the corresponding acronym will rarely have the accent present. Why, for example, does the Canadian federal government⁸ use "Ê" in "Ministère de l'Énergie, des Mines et des Ressources" but not in the corresponding acronym "EMR"?

Technology plays an integral role in the use of accented capitals and the alphabetization of accented words. The ability (or inability) of typewriters and typesetting machines to generate these letters has probably set today's accentuation trends by default. Similarly, today's computer hardware manufacturers, and developers of word processors, desktop publishing systems, and communications protocols, unknowingly influence current and future practices as they define their product offerings. To the extent that they provide easy processing for accented forms, they will encourage rather than discourage the wider use of accented letters.

For example, Canadian phone books seldom show the accent on the first letter of a person's last name; this is probably due to the fact that alphabetical sorting procedures, which operate on the name, would have to take these accented letters into account. Yet in the same phone books, accented word-initial capital letters are used when no alphabetic order is required, such as in the middle of a yellow-pages advertisement. Consider also the dictionary⁹ which had entries like "A BON MARCHÉ" in its 1952 edition--was "Â" avoided for technical reasons? This seems likely since a newer version of the same dictionary published in 1987 shows such entries fully accented. Here we see clues that technology may directly influence accent usage.
In the world of everyday writing, printing, and publishing, the use of accents is conditioned by habits, vague orthographic principles, personal preference, and even technology. While no issue of orthography needs to be fully legislated or documented, one senses that the characteristic French vigilance for proper linguistic form falls short in accentuation issues. This in turn directly impacts the field of natural language processing.

We will now examine another accent-related phenomenon which also has far-reaching implications for the French text processing industry.

DEACCENTED HOMOGRAPHS

A "homograph" is a word which has one lexical form, or spelling, but which is inherently ambiguous, representing several meanings. French homographs are plentiful; here are a few examples:

- est: inflection of the verb 'être' ("to be")
  noun meaning "east"

- noyer: verb meaning "to drown"
  noun meaning "nut tree"

- mises: 2 inflections of the verb 'miser' ("to bid")
  plural of the noun meaning "a positioning"
  1 inflection of the verb 'mettre' ("to put")

Many orthographic works have a good inventory of such homographs.

More interesting to us is the concept of "deaccented homographs". These are sets of words which are not true homographs, but do become homographs when accents are missing. For example, the words "jeune" and "jeûne" are not homographic when written as such, but are deaccented homographs when the circumflex is left off the latter.

In the headwords of the Petit Robert there are over 720 sets of deaccented homographs. Note that this number only includes base forms, but no inflections of any kind. There are thousands of sets of deaccented homographs in French when all inflected forms are taken into account.

The classification of such homographs reveals several different types. For example, most of the Robert headword deaccented homographs are of the type "noun/er-verb participle" such as:
  "abîme, abîmé"
  "bouche, bouché"
  "courbe, courbé"
"pénicille, pénicillé". These pairs are closely related semantically and share the same orthographic derivations. Formed from the most common verb conjugation paradigm, they constitute a very productive set.

Another kind of deaccented homograph set is one where ambiguity is introduced by the addition of a prefix such as "dé-" or "re-":

"denier, dênier"
"recréer, récréer"
"repartir, répartir"
"reformer, réformer".

Yet another is caused by different spellings for the same word, based on pronunciation, and often occurs when words come into French from another language. Usage often vacillates between the type of accent on a particular letter:

"foêne, foène"
"prêle, prêle"
"poêle, poêle"

or with the presence/absence of accents:

"gable, gàble"
"têpidarium, tépidarium"
"rodeo, rodéo"

or even the shifting of accents from one letter to another: "taôisme, taoïsme".

Such problems have caused much difficulty in the past, and in 1976 a French government departmental order decreed that variations like:

"référendum, référendum"
"événemen, évênement"
"épitre, épître"
"crûment, crument"

would be tolerated as long as a confusing homograph would not result (as would be the case with "tâche, tache" or "dites, dites").

Another common set of unaccented homographs results from the logographic role of accents mentioned earlier. More examples of such pairs include:

"sur, sûr"
"mur, múr"

Verbal conjugations are particularly productive in the creation of other types of deaccented homographs:

"tue, tué"
"prises, prisés"
"rîmes(pl), rîmes"
Finally, there are many sets of deaccented homographs whose convergence can be best described as "accidental":

"mais, mais"
"mâcon, maçon"
"même, mêmé"

Obviously, these are but a few of the thousands of deaccented homographs in the French language. These words only differ by accent information—sometimes with a complete change in meaning, sometimes with none. Accents, then, are often crucial to the resolution of homographs. It is when these accents are missing from text that computer-based language processing becomes problematic.

THE PROBLEM

As we have seen, it is not uncommon for words to be deaccented, especially when there are capital letters involved. Most people who know French are not seriously hindered by the absence of accents in a French text. Context can usually serve to eliminate any confusion that may arise. For example, one would not have any difficulty understanding:

Il travaille aux Etats-Unis.
Voilà l’Ecole militaire.
A Paris il neige très peu.

We can reason, upon seeing a sign saying SALLE DES CONGRES, that it probably announces a convention center, and not an eel warehouse. In some cases, though, even with our capacities for reasoning and processing contextual clues, a misunderstanding may still result. Both readings of each of these newspaper headlines are plausible:

UN POLICIER TUE
AUGMENTATION DES RETRAITES

Even whole sentences may not give enough information to a human for correct understanding of deaccented text. Consider:

ELLE EST A LA POURSUITE D’UN DIPLOME.

LES PREMIERS COLONS D’AMERIQUE S’ADAPTERENT VITE A LA CONSOMMATION DU MAIS.

IL Y A CETTE ANNEE UNE RECOLTE SURE POUR LE VIGNOBLE BORDELAIS.
Clearly the accent is sometimes necessary, even to the human reader:

(DIPLÔME or DIPLÔMÉ?)

(COLONs or CÔLONS?)

(SURE or SÛRE?).

Whereas the human may have little difficulty solving problems such as these, the computer is definitely at a disadvantage. The expertise and experience of human understanding far outstrip computational capacity. Yet computers are increasingly being called upon to handle such "fuzzy" and "grey" areas of accent usage as missing or out-of-order accents.

Special algorithms and methods must be developed to allow computers to "reaccent" deaccented text when necessary. Many different types of natural language processing software require this capability. We will now mention some of these applications.

The correct use of accents is problematic even for Francophones. Entering accented characters with word processors is often awkward—requiring several keystrokes—and therefore error-prone. Still, many word processors do not yet allow for accent error detection and correction. Some do not even permit accented capital letters; others do, but flag such words as erroneous.

Some text formatters have the ability to split words into hyphenated syllables. As with word processors, they can be very unforgiving programs when run against text not exhibiting the accentuation characteristics expected by the developers.

There is an increasing number of text-to-phonetic-representation and text-to-speech utility programs. The phonographic accents discussed earlier must be present in the input text to assure correct acoustic rendition of certain words.

The storage, management, and retrieval of terminology also requires the consideration of accent information. The resolution of deaccented homographs and accented alphabetization must be central to the design of a useable terminology management system.

Some natural language processing applications do varying degrees of morphological, syntactic, and even semantic parsing of the input text. Such programs are used for database querying, natural language translation, text
summarization, grammar and stylistics checking, and intelligence gathering. They, too, require accented input or the ability to reaccent input.

What is the extent of the problem? While it is true that most texts have accents in them, some texts may have suspect or missing accent information. Often TELEX messages, for example, are in all upper case and have no accents. Technology must compensate in such cases. In French technical text only 3% to 6% of all letters are accented. When full reaccentuation is required, though, the task is more demanding: a full 45% of all letters are the five vowels and "c", which can all potentially take accents. Every word is suspect in such an environment.

THE SOLUTIONS

There are several approaches to the reaccentuation of text. The following is by no means an inventory of all possible algorithms, but rather the ones I have successfully implemented.

The most obvious approach is a straightforward one: just calculate all of the word's possible accented permutations, and sift through or filter out the extraneous ones. Such an approach done on the word "plat" would produce the three possibilities: "plat", "plât", and "plàt". The last two could be discarded. This approach, though, quickly multiplies the number of possibilities. For example, consider these words and the number of permuted reaccentuation possibilities each has:

- **plat:** 3
- **eau:** 60
- **événement:** 625
- **élémentaire:** 5625
- **éventuellement:** 12,500
- **hétérogénéité:** 18,750.

This factorial explosion, or overgeneration, of possible forms makes the permutations approach, despite its simplicity, impractical.

One could constrain the number of results from this approach by adding rules such as:
- "e" cannot have a circumflex or grave at the end of a word
- "ê" cannot begin a word.

However, the use of such ad-hoc rules introduces a whole level of complexity and data management issues to the solution. The set of such rules could easily become open-ended, unintuitive, and internally conflicting.
A "cut-and-paste" morphology approach could be used where morphemes are identified, affixes stripped, and forms reaccented. For example, one could use rules like:
- if a word ends in "-atre", try it with "-âtre"
- if a word begins with "de-", try it with "dé-
- if a word begins with "aero-", try it with "aéro-".

Again, these rules could potentially become unmanageable due to their complexity and internal interference.

Some word processors have a list of all possible inflected words in French, usually at least 1,750,000 words, in their spelling dictionaries. The creation and updating of such a list is a very large undertaking indeed, as French is very rich in inflected forms. For words (and unaccented variants) not contained in the list, a straightforward word-list retrieval system cannot provide reasonable reaccented alternatives.

Another approach is one which I call "tiling n-graphs". It involves the use of n-graphs, or sequences of n letters. A tetragraph, for example, is a sequence of four letters. To "tile" n-graphs is to overlay them, much as tiles overlap each other on the roof of a house. We will work through an example of reaccentuation by tiling tetragraphs.

First, though, we identify and deaccent all tetragraphs in the French language. Mathematically, there is a possibility for well over 2.5 million sequences of four French letters (including one character for word boundaries). In practice, however, we only end up with some 7,300 deaccented tetragraphs in French which can take accents. Each deaccented tetragraph has a few reaccentuation possibilities, or tetragraph substitutions. Any tetragraph not in the dictionary cannot take an accent.

Here are a few examples of entries from the French tetragraph dictionary: (# is the word-boundary character)

<table>
<thead>
<tr>
<th>acad</th>
<th>#gai</th>
<th>depe</th>
<th>aine</th>
<th>gue#</th>
</tr>
</thead>
<tbody>
<tr>
<td>açad</td>
<td>#gai</td>
<td>dépe</td>
<td>aînê</td>
<td>guê</td>
</tr>
<tr>
<td>açad</td>
<td>#gai</td>
<td>dépe</td>
<td>aînê</td>
<td>guê</td>
</tr>
<tr>
<td>açad</td>
<td>#gai</td>
<td>dépe</td>
<td>aînê</td>
<td>guê</td>
</tr>
</tbody>
</table>

This is how tetragraph tiling is done:

1) the input word (with its boundaries) is "exploded" into tetragraphs
2) each successive tetragraph is looked up in the tetragraph dictionary

3) successful tetragraph substitutions are saved and carried on to the next iteration, while those which fail because of contradiction to pre-existing conditions are discarded.

Here is an example showing: the source word "#evenement#"; its tetragraphs and reaccented variants supplied from the deaccented tetragraph dictionary; and iterated matching and discarding of reaccented variant strings. The single reaccented form #événement# is the final result. {} means the entry is not in the dictionary--use the tetragraph exactly as it appears, with no reaccented variants possible.

#evenement#

#éve even vene enem neme emen ment ent#

#éve even vene énem neme emen {}
#évè évèn vene ènem nème émen
#évè évèn vene nème émèn
#éven vènè vènè nèmè èmèn
#évene
#événe
#événen
#événenem
#événenement
#événenement#

As can be seen, in such an example as this one, we have avoided overgeneration by applying data on the immediate context of French accented letters. In fact, the tetragraphs capture inherently the immediate lexical environment of
accented letters. Using this data, any new word can be successfully analyzed.

Here are the numbers of reaccented possibilities supplied by tiling tetragraphs on the words shown in the permutations examples:

<table>
<thead>
<tr>
<th>Word</th>
<th>Possibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>plat</td>
<td>1</td>
</tr>
<tr>
<td>eau</td>
<td>1</td>
</tr>
<tr>
<td>événement</td>
<td>1</td>
</tr>
<tr>
<td>élémentaire</td>
<td>4</td>
</tr>
<tr>
<td>éventuellement</td>
<td>1</td>
</tr>
<tr>
<td>hétérogénéité</td>
<td>2</td>
</tr>
</tbody>
</table>

This tiling approach can be used with digraphs, trigraphs, or any other length of n-graphs. For our particular implementation's memory and disk storage requirements, processing with tetragraphs proved to be the best approach. Dynamic updating of the n-graph dictionary can be easily accomplished automatically as new forms are found while handling input texts. The tetragraphs and their data can be stored in an indexed file, hash table, lexical trie, or finite-state automaton.

This tiling n-graph approach, while avoiding overgeneration to a great degree, nonetheless does generate incorrect forms. For example, consider these 4 forms generated for "élémentaire":

élémentaire, élementairé, élementaire, élementaire.

One way to constrain the generation of such forms is to assign to each tetragraph substitution a preference or weight based on its frequency of occurrence in French. If we consider the forms mentioned above, and redo the tiling with weighted substitution tetragraphs, we reject all incorrect reaccented forms for "élémentaire" and "hétérogénéité" by a wide margin.

One other reaccentuation approach involves the lexicon. Natural language processing applications usually consult an on-line dictionary or lexicon. While containing application-specific data, this same lexicon can also provide reaccentuation data for known words. In lexicons storing entries in alphabetic order, keyword letters and their accents can be considered separately, rather than together. Of primary consideration is the letter; the accent, though important, is secondary. The lexicon's data can then be sequentially organized, primarily by the entries, and if necessary based on any accents. Deaccented homographs can then be stored contiguously in the lexicon. Once a direct read into the lexicon has been performed, either successfully or unsuccessfully, comparatively inexpensive sequential reads can be performed to collect any
or all deaccented homographs. Any accent information on the original word can be consulted to filter out inappropriate results.

USER INTERFACE

All of the above approaches have been implemented in practical systems with varying degrees of success. Ultimately, though, they must fit the requirements of a human user. The user’s needs can best be met when he is allowed to select the degree of reaccentuation needed. Reaccenting requires noticeable overhead, so it is in the user’s best interest to select a level of processing best suited to the source text’s characteristics.

The user can control the system based on letter-case accenting considerations. REACCENT NOTHING means that the input text is assumed to be correctly accented. REACCENT CAPITALS ONLY means that the input text has deaccented capital letters, so the system will automatically reaccent capitals. REACCENT ALL LETTERS means that the input text is missing accents on both lowercase and uppercase letters. Any accents present will be accepted as valid, but any unaccented letters are reaccented when possible.

The user can also control processing based upon the reliability of the input text itself. TRUST ACCENTS means that the accent information is expected to be there, and also to be correct. TRUST OVERT ACCENTS means that if a word has an accent anywhere on it, it is safe to assume that all accent information present for that word is correct. TRUST NOTHING means that no accent information (i.e. the presence or absence of accents on each letter) is considered reliable, and all accents are ignored.

SUMMARY

Although French accentuation is a complex orthographic problem complicated by inconsistent usage and constrained by technology, there are ways to minimize its impact on French natural language processing applications by providing for user-controlled reaccentuation of texts deficient in accents.
FOOTNOTE REFERENCES


6. Aurel Ramat, Grammaire typographique (Québec: 1982), pp. 65 and 70.


BIBLIOGRAPHY


Orwell's Language and Thought in
"Politics and the English Language" and 1984

Don Chapman
Brigham Young University

When I first started this paper, I expected to find near-unanimous approval of Orwell's insight on language. After all, "Politics and the English Language" has been widely anthologized, while "double-think" and "Newspeak" have become common terms. In my composition classes, I had heard his name treated with the same respect I associate with Chomsky and other lofty figures in language thought, so by association, I regarded Orwell highly important, if not revolutionary. I expected most people would agree with my anthology:

- Orwell had explored in one of his most influential essays, "Politics and the English Language," the decay of language and the ways in which it might be checked. The forty years that have passed since it was written have only confirmed the accuracy of its diagnosis and the value of its prescription (Norton 2260).

To be sure, Orwell may be one of the most capable prose writers ever, and he seems to speak sense to us. In a time when missiles are called "peace keepers" and taxes "revenue enhancements" Orwell's objections to dishonest political writing seem relevant and incisive. But regardless of Orwell's popular reputation (and I'm not sure he really sought for such a reputation), few linguists share such veneration. As Quirk claims, "the time [is] right for a reevaluation of Orwell, especially in view of the undue reverence in which he is held as a serious thinker on social and linguistic matters" (48). I'm afraid I agree with Quirk, since Orwell's ideas are generally unoriginal and lack a sound theoretical foundation.

We might begin with his status as a "revolutionary" thinker, when in fact, his thoughts on language are highly derivative. Quirk points out, the notion of language as a "Loaded Weapon," a manipulative tool, has been around for centuries. Francis Bacon had spoken of people "too ready to be moved by words themselves without thought to what weight of matter they connoted" (49). And later Goethe's Mephistopheles says:

- where concepts fail,
- At the right time a word is thrust in there.
- With words we fitly can our foes assail,
- With words a system we prepare,
- Words we quite fitly can believe (Faust I.1900-04).

During the 20's and 30's, the chilling harangues of Hitler, Stalin, and others heightened fears of language manipulation and gave new impetus to the plain language movement which sought to combat such manipulation. These were the decades of the Fowlers, A.P. Herbert, Eric Partridge, and Ivor Brown. Orwell was drawn heavily to this movement, and often included its ideas in his writing. In
"Politics and the English Language," for example, Orwell clearly links decayed language with degenerate politics: "Modern English . . . is full of bad habits . . . If one gets rid of these habits one can think more clearly, and to think clearly is a necessary first step toward political regeneration" (128). And in his summary, Orwell claims "If you simplify your English, you are freed from the worst follies of orthodoxy" (139).

Quirk further claims that "Politics" is "little more than an expansion of the five maxims set forth on the first page of the King's English by the Fowler brothers in 1906" (50). A quick comparison between the two illustrates Quirk's point:

**ORWELL**

1. Never use a metaphor, simile, or other figure of speech which you are used to seeing in print.

2. Never use a long word where a short one will do.

3. If it is possible to cut a word out, always cut it out.

4. Never use the passive where you can use the active.

5. Never use a foreign phrase, a scientific word, or a jargon word if you can think of an everyday English equivalent.

6. Break any of these rules sooner than say anything outright barbarous.

**FOWLER**

- Prefer the short word to the long.
- Prefer the single word to the circumduction.
- Prefer the familiar word to the far-fetched.
- Prefer the Saxon word to the Romance.
- Prefer the concrete word to the abstract.

As further evidence that Orwell was mainly repeating the thought of his day, Bolton notes essential resemblances between "Politics" and The Art of Writing by Sir Arthur Quiller-Couch.
Quiller-Couch:
So long as you prefer abstract words, which express other men’s summarised concepts of things, to concrete ones which lie as near as can be reached to things themselves and are the first-hand material for your thoughts, you will remain, at the best, writers at second-hand (qtd. in Bolton 192).

Orwell:
When you think of something abstract you are inclined to use words from the start, and unless you make a conscious effort to prevent it, the existing dialect will come rushing in and do the job for you, at the expense of blurring or even changing you meaning (138).

Yet I don’t want to criticize Orwell too harshly for being unoriginal. After all, my own paper is highly derivative, owing much to Bolton, Quirk, and others. The more serious charge is Orwell’s shaky, perhaps naive theoretical underpinnings. Fundamentally, Orwell seems to view language as an object, something separate from ourselves. This view manifests itself throughout "Politics," in Orwell’s unsound notions that language can be corrupted or engineered, and that a language controls thought and vice versa.

Let’s begin with his assertion that language can be corrupted. Orwell makes his position clear from the first sentence of "Politics": "Most people who bother with the matter at all would admit that the English language is in a bad way" (127). Later he states "But if thought corrupts language, language can also corrupt thought" (137). His essay is full of words like "decay," "corrupt," and "decline." Such claims illustrate how Orwell reifies language, since they invite an analogy to substantial objects. Presumably, language can be debased, just as gold is debased by copper, or pure water is contaminated by oil spills.

But how does one measure "decay"? Few, if any, trained linguists would claim competency to measure "corruption." Of course, Orwell never rigorously defines what he means by corruption—presumably he assumes words like "corrupt," "decline," and "deteriorate" are self-evident. Yet speaking of language corruption presumes some standard from which the language has degenerated, and such a view naively overlooks the history of language. After all, language has changed continuously, so where do we find the standard? Is Shakespeare’s English corrupt because it differs from Chaucer’s? Is Dryden’s English more corrupt than Shakespeare’s? And what about all the languages descended from Indo-European? Are they all corrupt? Or in Bolton’s terms, "is Spanish merely corrupt English?" (33). Certainly Orwell was right to confess he doesn’t have "sufficient knowledge to verify that the German, Russian and Italian languages have all deteriorated in the last ten or fifteen years, as a result of dictatorship" (137), but he shouldn’t have been any more confident about measuring
deterioration in English, either.

We might look to the contents of the essay—the "catalogue of swindles and perversions,"—to find what Orwell means by corruption, but unfortunately we only find complaints about current usages in English, not the language itself. "Staleness of imagery" and "vagueness of expression" reflect more on particular speakers of English than on English itself. He hasn't proved at all, and I doubt he can, that English lacks the resources to express ideas precisely; to do so he would have to show that no English speaker, including himself, is capable of precise expression. We can't really assent to such a position.

Orwell's materialist bent manifests itself again when he maintains language can be engineered or consciously changed and crafted for our purposes. Early on he speaks of language as "an instrument which we shape for our own purposes" (127), and later he explicitly claims: "the decadence of our language is probably curable. . . . Silly words and expressions have often disappeared, not through any evolutionary process but owing to the conscious action of a minority" (137-38). Of course Orwell's attitudes on language doctoring receive their most celebrated form in Newspeak, the language of 1984 in which words are stripped of secondary meanings, so that unorthodox expression is impossible.

But Orwell shouldn't be too confident of consciously-crafted change in language, since previous attempts to engineer language have largely failed. Admittedly, sexist language, such as "postman" and the generic "he," have generally been excluded from published writing, but perhaps we ought to regard such changes as conventions enforced by editors, like punctuation rules, since most people are still saying "salesman," regardless of what appears in print. Furthermore, Orwell's own examples undercut him, since "explore every avenue" and "leave no stone unturned" have certainly not been laughed out of existence as he claims (138). In fact, these stubborn phrases are much more common than several of the hackneyed phrases he still wants to eliminate, such as "iron heel," and "bloodstained tyranny."

Finally, Orwell treads on shaky ground again when he deals with the relationship between language and thought. Characteristically, he displays much more confidence than most who write about this complex relationship, as he writes his opinions boldly, even though no consensus has been reached. Essentially he espouses two contradictory views of the relationship between thought and language. On one hand he assumes we can think without language:

> When you think of a concrete object, you think wordlessly, and then if you want to describe the thing you have been visualizing you probably hunt about till you find the exact words that seem to fit it. . . . Probably it is better to put off using words as long as possible and get one's meaning as clear as once an through pictures or sensations (138).

At another point, Orwell speaks of a writer who "is not seeing a mental image
of the objects he is naming; in other words he is not really thinking" (134).

On the other hand Orwell assumes language can take over the thought process entirely:

When you think of something abstract you are more inclined to use words from the start, and unless you make a conscious effort to prevent it, the existing dialect will come rushing in and do the job for you, at the expense of blurring or even changing your meaning (138).

Elsewhere he write of "throwing your mind open and letting the ready-made phrases come crowding in" (135), and "appropriate noises coming out of [the] larynx, but [the] brain is not involved as it would be if [the speaker] were choosing his words for himself" (136).

So on one hand, thought creates words, on the other words create thought. How can both views be valid? But perhaps neither is, since each presents problems. Let's look more carefully at them, again emphasizing that most views on language and thought remain quite speculative.

Characteristically, Orwell fails to define language or thought. For someone complaining about imprecision, he leaves a lot of loose ends. But by inference, it appears he believes that thinking can be entirely separate from language. Somehow, we can apprehend the world before we speak. He describes a prelinguistic experience, in which we comprehend the world, presumably in images or sensations.

Now this seems sensical enough, after all, we've all experienced those moments in which we know (or at least think we know) what we want to say, but can't put it into words. Other times, our words don't seem to say what we really mean. But can this really be explained by saying we have a clear idea in our heads before we try to speak? Wicker disparages such a view, claiming it "insists that a prelinguistic experience of undifferentiated sensations gives an immediate knowledge of how things are, and so provides the basis for all certainties" (qtd. in Bolton 34-35). Much current philosophy rejects such a strong separation between language and thought. As Palmer explains,

Since we categorise the objects of our experience with the aid of language, it may be the case that learning about the world and learning about language are activities that can not be separated and that therefore our world is partly determined by our language (44).

Part of the problem again is Orwell's insistence on language as an object existing outside of the mind, instead of a system within the mind by which we interpret the world. In accordance with his materialist bent, he implies that language can and should correspond to something in the "real" world. For example, he writes

Words like romantic, plastic, values, human, dead, sentimental, natural,
vitality, as used in art criticism, are strictly meaningless, in the sense that they not only do not point to any discoverable object, but are hardly ever expected to do so by the reader (132).

Using language, then, consists of finding names for conceptions we've already arrived at.

It's not likely, however, that language works through such simple correspondences. Current views, such as Sausurrean semiotics, reject external references. (Bolton points out fairly that while Sausurre's lectures were published in 1916, they weren't translated into English until 1959. It's understandable that Orwell wouldn't have been influenced by them (38)). Sausurre interprets language as a closed system in which a signified (sound or written symbol) corresponds to a signifier (concept within a system, not at all tied to an external referent). The language system is entirely autonomous, and allows us, as Bolton points out, to say such things as "a four-sided triangle," even though we'll never find one in the external world (Bolton 35).

The other assumption Orwell implies is that language can take over the thought process. At his extreme, Orwell says "ready made phrases come crowding in [to your mind]. They will construct your sentences for you--even think your thoughts for you, to a certain extent." Although he qualifies his remark with "to a certain extent," his other remarks clearly show that he regards that certain extent as extensive indeed--that words almost entirely take up our thought process. Again this seems sensical--at least we have known those who parrot words without really paying attention to what they are saying. But does that mean the words are doing the thinking for them? I think what he really means is that we pass on a phrase we have heard without exploring its underlying implications and assumptions, without investigating its details. Were we to do so, we might not agree with the phrase. In the sense that we don't pursue the possibilities suggested by a phrase we aren't extending our thinking, but we can't say the words are doing the thinking for us. Again, the problem isn't with language, but with lazy speakers of the language.

In conclusion, I don't mean to undermine Orwell's contributions. He seems to speak sense to us, and we can probably benefit from his exhortations to be conscious of language, both our own and other's. But we ought to be careful not to make Orwell into a language expert. Instead let's remember that he was primarily a journalist who wrote clearly what he observed.
Notes

1 I realize it's hard to tell just what Orwell meant to do with Newspeak--whether it is a satire on Ogden's Basic English, or whether it expresses Orwell's earnest beliefs about trends in language. But given his explicit statements in "Politics," it's clear Orwell thought such institutional language doctoring possible.
Works Cited


The following texts are the work of a so-called translator, published by a client who believed that he got precisely what he was looking for to communicate his message.

"A las puentes naturales monumento nacional de cuentos piedras sólidas alcanza a través cañones profundos formando las puestas más grandes naturales en ninguna parte. Y entonces hay el más grande: La puente Arco iris monumento Nacional. Su tamaño interior es tan alto (309 pies/94m) que el capital de la nación lo podría encajarse de alaño. En el lugar de los cuatro esquinas es el Nacional de Hovenweep Monumento, seis grupos de tores graneros y viviendos abandonados misteriosamente hace cien años por los ancestrors de los indios pueblos."

"Ind every space we find synpathelic places in every color as Purhuay, Ahuaycha, Pamuri and others, were wits a little art we can feed a delicious popular alcoholic beverage of the style of mother Etelvina.

This Pampas with his shinning sky annoys blue with love received airs with his wide highway to Huancayo, Ayacucho, Cobriza, with his big typie January foty of 576 hours dancing without stop with many dancers in moon’s nights, in the best time with the best time teacher a guitar Direct "Chusco" Gamarra, well leave back the Pamp with his farh full of redish and enchantament of their streets, and soy good by of the many and nice fusids, who made durinshortstay remember the swahlows and music nigths in Snack Bar "103" where we receive a perfectly attention."

When viewing samples such as these, people tend to react in several ways: some are moved to laughter, others feel that they would never do such a poor job, and still others think that these samples are the exception rather than the norm. Unfortunately, this is not a laughing matter; most people who do not know what translation is all about, do
produce texts similar to these, and furthermore, this problem is quite prevalent in the nation.

Given these facts, let's take a look at some of the things every would-be translator should know about translation.

The life of a translator revolves around theory and practice. By theory we mean the set of rules and regulations that govern the process known as translation. By practice we mean the protocol that regulates the interaction between client and translator. For the purpose of this presentation, we will discuss the basic concepts relating to theory and practice that every would-be translator should know before he or she attempts to work as a translator.

First of all, let us define translation. Translation is a process by which a text originally produced in one language is rendered into another, and the finished product is in written form. This is an important point: every would-be translator should know that translation involves a written process, as opposed to interpretation, which is the oral transfer of a text from one language into another. Furthermore, a translation can only be called a translation when the target text reproduces the same reaction, the same information and the same impression that the original did on its intended audience. Anything short of this cannot be called a translation.

Second, every translation process presupposes an interaction between meaning and form. A translator must transfer the meaning - i.e., the information, the message - contained in the source text, and he must do it accurately. What most people don't understand is that in order to have a good translation, the information must be rendered in the form of the target language; that is to say, using the syntactic patterns typical of that language. The source text provides the information; the translator must provide the appropriate target language form to make that information available to his target audience. If he fails to do that, if he allows the form of the source text to show up in his target language, then he has created what Alan Duff calls "a third language"; that is to say, a text similar to the examples presented in this paper.

Third, if form is so important in a translation, then it stands to reason that a translator must translate into his native language. This is the form with which he or she is most familiar; this is the language he has grown up with. The examples used at the beginning of this presentation also illustrate what happens when a translator attempts to translate into a foreign language rather than into his own.

This brings us to our fourth point: what every would-be translator should know about his own limitations. Contrary to what many people believe, there is more to translation than the language skills developed in a foreign language program. To be a competent translator, one must be an expert in his own native language and have an almost equal command of the foreign language. Being able to function in a foreign language is not the same as being able to produce a translation that sounds as if it had been written in that language from the very
beginning. And yet, it is not a folly to state that every person who took a foreign language course at one time or another felt that he could translate, not only into his own native language but into the foreign tongue.

Fifth, every would-be translator should know the four basic steps of the translation process: 1. Text analysis, 2. Documentation, 3. Translation, and 4. Revision. By text analysis we mean reading the entire text before attempting to translate it. Such reading should provide the would-be translator with information regarding meaning, form, and problems to be solved, as well as the tone and purpose of the text, and should also bring out information concerning its intended target audience. To translate without doing a text analysis is foolish.

Once the text has been analyzed, the translator does not begin his translation but rather goes about collecting all the materials necessary to resolve the problems he has discovered in the text. By documentation we mean the gathering of all tools (glossaries, dictionaries, other similar texts already translated, colleagues, encyclopedias dealing with that particular subject matter, grammar books, illustrated dictionaries, etc.) from which the translator attempts to develop a terminology file to use with that text. Nowadays, these files are produced with the aid of the computer and by using a text data manager such as Lexitem, among others.

The first two steps of every translation project -text analysis and documentation- require a great deal of time and effort, and many would-be translators are not aware of how important it is to spend that time in preparation for a translation.

The third step is the actual translation. In this day and age, no one handwrites a translation. Most translators use a word processor with a built in terminology file. This terminology file is produced during the documentation stage of the project, and it may undergo some modification as it interacts with the text being produced during the translation process, but at the same time it will grow with each translation project, providing, therefore, a wealth of information for future reference.

The translation thus produced, however, is rarely the finished product. It will go through many more drafts until meaning and form merge into one cohesive discourse that no longer resembles a text written in another language.

Then, it is ready for the final stage, which is called revision. Every would-be translator should know that this is a vital part of every translation, and that there is a system to it. First, no revision should be undertaken immediately after completing the translation. The ideal situation calls for a 2 week interval between the completion of the translation and the revision process, if the revisor is going to be the translator himself. By then, he can read his translation and spot "third language" problems more readily than if the source text were still fresh in his mind. Once he has checked the form, he can proceed
to check the meaning. The real revisor, however, should be someone who
has not read the source text and whose native language is the language
of the translation. This brings us to another point. Is there more
than one good translation? The answer is Yes. If a translation is an
interaction between meaning and form, different translator may chose
different forms to convey the same meaning.

From all of this we may derive that what a would-be translator should
know about the theoretical aspect of translation is that there is more
to it than language knowledge, and that a good dictionary alone does
not make a good translator.

Let us say, however, that the would-be translator has learned the
theory and is ready to go to work. What should he know about the
practice of translation?

First of all, he must decide whether he wants to free-lance or work in-
house. Free-lancers have the option to work for themselves, or for an
agency on a part-time basis. In-house translators may work in the
translation department of a company or at a translation agency.
Regardless of where he works, however, he must be aware that
translators need to specialize, and that no one should attempt to
translate anything and everything.

Second, more and more clients are aware that there is more to
translation than language skills, and therefore they may request proof
that the would-be translator is competent. One such proof is a degree
from a reputable institution specialized in the training of
translators, and the other is national accreditation granted by a
professional organization. At this time the only such organization is
the American Translators Association. Third, a would-be translator
should know something about what is known as translator/client
relations. In most cases, the client knows practically nothing
concerning the translation process. In his mind, if a person has
studied a foreign language and has a dictionary, he ought to be able to
translate. It is only in cases where the client has been "burned" by a
poor job, that he is willing to abide by the rules of the profession.
Therefore, it is up to the translator to educate his client as to why
he had to hire a professional translator, and what he is paying for.
In view of that, a would-be translator needs to be up to date on how
much to charge his client and how to justify that charge. In the case
of a free-lancer, he must also know how to protect himself against
unscrupulous clients. Matters such as contracts, retainer fees, hidden
costs, pre-payment, deadlines, and the likes, should be clearly
understood and presented to the client in a professional bid.
Furthermore, the translator should know how to make his translations
available in camera ready form, or via modem or FAX. He should also
make provisions to retain control over any changes made in his text.

If the translator is working in-house, whether for an agency or a
company, his job may be determined by the rules of the house. In that
case, a would-be translator would do well if he would ask for a set of
the rules -in writing- that he must uphold while working for the firm.
Some firms, for example, hire translators as free-lancers, but they do
not allow them to work for other clients. A would-be translator should also know that, in most cases, an agency will charge a client at least triple what he himself is being paid for his work. The rest of the cost goes for over-head expenses.

Lastly, a would-be translator should know that there is always somebody willing to work for less, and that sadly enough, a large percentage of clients will go for the cheaper bid. In such cases, it is not unusual to find that the would-be client does come back later with the finished product and asks you to correct it, because someone who speaks the target language has warned him that, in his opinion, the translation is not quite right. At times such as these, a translator often finds out that as long as there are so called "would-be translators" out there, there will be work for him to do.
INTRODUCTION

For the past 4 years I have been in the Northern Territory of Australia working on a number of projects involving the Australian Aboriginal people of the region. The experience has been very rewarding and one which has taught me much about cross-cultural communication and misunderstandings that occur in such an environment.

Some of what I have learnt can be summarized in one anecdote involving a linguist who had been travelling around the "Top End" conducting field work with Aboriginal languages. It was a hot and sultry day, as all days are in that area of Australia. He was near the ocean and eventually came across an inlet that looked perfect for a cool down swim. Naturally he was concerned about crocodiles. Noticing a small group of Aboriginal men standing near the inlet, he approached them and said in his best Aboriginal creole, "Goodday, jeya aligerra?" meaning, "Is there a crocodile in there?" "Na, jeya na aligerra, na crocodile, boss," was the reply. Relieved, he jumped into the water and swam around for quite a while. Eventually he got out and started talking with the Aboriginal men who had been lingering nearby. He asked them why they hadn’t gone for a swim and received the reply, "Na boss, plenty big mob, bigpela shark jeya boss."

The linguist had studied the Aboriginal languages and creole at the sentence level, but had not acquired the discourse rules of these languages. If he had acquired these rules, he would have known that when one approaches an Aboriginal seeking information he must first establish some form of geographical or personal commonality with the Aboriginal such as someone that they both know or a place where they had both been. One usually finds something in common. Once that is established, meaningful communication can commence. If this vital discourse rule is absent, then the Aboriginal may not answer the question, or will answer it with the barest of information as in the case above.

As this anecdote suggests, the sociocultural distance between Aboriginal Australians and other Australians is significant. Consequently, the process of acquiring standard English
communicative competence by the Australian Aboriginal people, particularly those living in the more isolated areas of the Northern Territory of Australia, provides a rich environment where areas of learner difficulty become highlighted. One such area of difficulty involves the "rules" or procedures to be followed if one is to operate at full communicative competence at and beyond the sentence level in written discourse.

Sufficient evidence is available to indicate that different speech act rules exist for different languages with these specific rules related to the cultural and sociolinguistic dimension within a particular speech community. Kaplan (1966; 1972) has demonstrated that discourse rules form patterns which can be related to cultural systems and that these patterns are evident when written texts are examined. Consequently, while numerous forms of developing meaning are available to all languages, each language exhibits clear preferences as to the presentation of that meaning. Thus, as Kaplan (1987) states, there are:

important differences between languages in the way in which discourse topic is identified in a text and the way which discourse topic is developed in terms of exemplification, definition and so on (p. 10).

Or, as Clyne (1985) suggests:

it is the cultural value system that determines whether, to a particular group, directness is vulgar or indirectness is devious . . . whether a letter should come to the point immediately or gradually build up to the central speech act . . . whether linearity in discourse is seen as the only logical or comprehensible structure, or whether it is felt to curb exhaustive discussion (p. 14).

The underlying hypothesis of this present paper is that one aspect of these language and cultural differences, namely the difference between the culturally influenced discourse styles of Aboriginal writer and native English speaking reader, and vice versa, significantly impairs the educational attainment of Aboriginal child and adult students. In other words, Aboriginal people tend not to develop meaning in the same manner as English speakers prefer to develop meaning. Hopefully, some of the conclusions from this study can be applied to all contexts where writers from one language and culture are required to write in the discourse patterns of a second language and culture.

THE NORTHERN TERRITORY ABORIGINAL SPEECH COMMUNITY

Before commencing with the study, it is probably necessary to
provide a brief description of the Northern Territory Aboriginal speech community. An estimated 300,000 indigenous people inhabited the Australian continent at the time of initial European settlement in 1788. These people spoke between 200 and 650 languages, depending upon which definitions of the terms ‘language’ and ‘dialect’ are used\(^1\) (Senate Standing Committee 1984). The two hundred years since European settlement have seen a dramatic decline in these languages to the point where only eight languages survive today with more than 1000 speakers (Baldauf and Eggington 1989), five of these languages are in the Northern Territory of Australia. Black (1983) estimates that, in addition to these five languages, there are twenty-five languages surviving in the Northern Territory with one hundred or more speakers. Thus, of the 35,000 Aboriginal people living in the Northern Territory, there are 20,000 speakers of one or more Aboriginal language.

In the early days of European settlement in the Northern Territory an English based contact language developed which has followed the contact language, minimal pidgin, pidgin, extended pidgin, initial creole to extended creole continuum (Todd 1974; Muhlhausler 1974, 1986; Romaine 1988). This creole or ‘Kriol’ is becoming the lingua franca of the Aboriginal people of the Northern Territory with an estimated 20,000 speakers (Sandefur and Harris 1986).

However, the English language remains as the dominant language in almost all domains requiring interaction with the ever-present non-Aboriginal society. It is the language of communication with government, health, commerce and education programs. Unfortunately, significant communication barriers exist due to a number of factors including inadequate English language proficiency levels among the Aboriginal people, cultural insensitivity among the English speaking non-Aboriginal people and huge differences in communication strategies between the two groups (Shimpo 1985). The English spoken by the Aboriginal people, Aboriginal English, exists as a non-standard, low status variety of the language.

In an effort to better meet the educational needs of the Northern Territory Aboriginal people, the Australian government introduced a Bilingual Education program in 1972. This program now consists of 16 bilingual schools and has had mixed results in achieving its stated objectives (Eggington and Baldauf, 1989). In general, educational achievement levels in the Northern Territory for Aboriginal people are significantly below national standards and well below Aboriginal student standards in other Australian states.

\(^1\)See Chambers and Trudgill (1980) on difficulties faced when attempting to define the terms ‘language’ and ‘dialect’.
TOPOCAL DEVELOPMENT IN ENGLISH ACADEMIC WRITTEN TEXT

In order to set a basis for a discussion of difficulties Aboriginal writers may face when required to function in English, it is first necessary to briefly review topical development in English academic written text. Lautamatti (1987) has examined the relationship between discourse topic and sub-topics explaining the development of topic:

in terms of succession of hierarchically ordered sub-topics, each of which contributes to the discourse topic, and is treated as a sequence of ideas, expressed in the written language as sentences (1987:87).

Topical progression comes about generally through two types of sub-topic development:

1. parallel progression where the sub-topic in a series of sentences is the same, and

2. sequential progression where the topic of a sentence is provided by the predicate of the preceding sentence.

It appears that essential elements in the expectations of the reader of English academic prose are that there is a hierarchical progression of topic and that there is a "a direct and uninterrupted flow of information" (Kaplan 1987:10). Consequently parallel and sequential topical progression must add to the topic within a narrow set of parameters seldom, if ever, digressing from the stated, clearly defined topic.

ABORIGINAL LANGUAGE DISCOURSE PATTERNS

Aboriginal languages are oral languages and it has only been during the recent past that serious attempts have been made to encourage the development of literacy. Therefore, to commence this research into Aboriginal English discourse patterns, the nature of oral narrative styles in two Aboriginal languages as given in Texts 1 and 2 will be examined. Text 1 is a direct translation from Nunggubuyu, a language spoken at the Numbulwar Mission and was chosen from a collection of ethnographic material translated by Heath (1980). Text 2 is a direct translation from the Tiwi language of Melville and Bathurst Islands (Osborne 1974). Both these texts were selected at random from within their
respective collections. In these and other texts discussed in this paper, discourse units are numbered for reference.

Text 1 describes culturally determined brother-sister avoidance procedures. These avoidance relationships are a very important feature of traditional Aboriginal culture.

Text 1 (Adult, Oral Narrative in Nunggubuyu)

1That brother and his sister; that man should not go close. 2If his sisters are sitting somewhere together, 3he should not go close to them.

4That man will not go close to there. 5He will not stand nearby. 6He will stop far away 8and he will speak. 9He will ask them a question with words, 10but he will not get too close. 11That is how the Nunggubuyu behave (Heath 1980:342, Texts 77.1, 77.2).

The topic sentence (1) is developed through a series of parallel discourse units which are either repetitions, with slight variation, of the head topic "should not go close to them" (3, 4, 10) or synonymous phrases (5, 6, 7). Thus only four discourse units (2, 8, 9, 11) develop the topic further than given in the topic sentence. 2

Text 2 describes the origins of the crocodile.

Text 2 (Adult, Oral Narrative in Tiwi)

1I am going to talk about Jerekepai (crocodile).

2Long ago when he was a man, 3he lived at Waiperali. 4He had many wives. 5They were cracking xamia palm nuts. 6He was making spears, spears that spear, spear. 7He was making them 8and his wives were cracking xamia palm nuts. 9The others were all making baskets.

10Some marauders crept up there. 11They took a look, and, 12'He is making spears', they said, 13'he is making spears.' 14They got ready, 15and 'Oh!' they shouted. 16He ran 17while he was making spears. 18We gave him that name because he ran 19while he was making spears. 20The sea! 21He went under, under, under, under, under, 22and then -- the spear came up first. 23'You are the crocodile now,' they said. 24Tajuni, takampunga, they called him, jerengkepetuni. 25They called him that 26because there are many crocodiles in the sea. 27In the

2The original text in Nunggubuyu repeats the word ana-warubaj (nearby) followed by the negative marker yagi five times and uses the word malanga-nYanay (far away) twice.
beginning he lived as a man (Osborne 1974:101-102).

It is obvious that the speaker is attempting to emphasize the relationship between Jerekepai, his spears and crocodiles. This emphasis is developed through either direct repetition of the phrase "he was/is making spears/them" (Discourse units 6, 7, 12, 13, 17, 19) or allusion to spears (22). Thus the main topic is developed through repetition. 3

From a body of data similar to the examples given above, it is possible to generalize that, in oral Aboriginal language narrative style, discourse topic tends to be developed through the repetition or synonymity of the head topic discourse unit, and that this repetition acts as a cohesive device unifying the text.

ABORIGINAL ENGLISH DISCOURSE PATTERNS

Text 3 was written in Aboriginal English and is taken from a booklet produced by the Northern Territory Department of Education containing samples of community (rural) Aboriginal children's writing in English.

Text 3 (Child, English)

1 We alway go fishing at Marrm every holiday. 2 And if you go fishing 3 you must carry food. 4 And I caught one fish. 5 When I go fishing 6 I alway put things ready 7 before I go fishing. 8 sometime we go swimming. 9 When me and my grandmother went for fishing at Marrm 10 she caught one turtle. 11 And we cook it. 12 And I like going fishing. 13 And some people caught one big fish. 14 And then we went home (Northern Territory Department of Education 1985:20).

The topic of "going" fishing is introduced (1) but then a digression is made involving preparation for the expedition (2,3). This is followed by a report on the catch (4); next a return to the preparation theme (5,6,7). The development of these themes is interrupted by a swimming digression (8), after which we are returned to the catch theme (9,10,11). An evaluation of the event is made (12), a further development of the catch theme (13) and then the text concludes through a return home closure (14).

It may be that this is an example of a developing loose chronological ordering style, but, from a native English speaker point of view, one cannot ignore the digressions and the lack of linear topical progression. However, the topic

3 The original Tiwi text mentions ju-wunti-kerem-ani arawuningkiri (he make spear) six times, and arawuningkiri (spear) an additional five times.
develops along similar lines as shown in Texts 1 and 2. The head topic "go fishing" is repeated five times (Discourse Units 1, 2, 5, 7, 12) with one synonymous reference (9). A sub-topic "caught one fish/turtle" is repeated three times (4, 10, 13). Consequently, it may be that the child writer is exhibiting those text developing procedures relying upon repetition and synonymy of the head topic inherent in her native language discourse patterns.

Note also the tendency to alternate sub-topics. Topic progression develops from "going fishing" (1, 2), preparation (2-3), catch report (4), go fishing (5), preparation (6), go fishing (7), swimming (8), went (go) fishing (9), catch report (10, 11), going fishing (12), closure (13).

Texts 4 and 5 were collected and analyzed by Richards (1985) and come from Aboriginal college students attending a teacher training institute at Batchelor in the Northern Territory. Most of these students have developed their English proficiency in an ESL context.

Text 4 (Adult ESL College)

Batchelor College is at the township of Batchelor. This is the place where people come from many different places to do teacher training. The places where people come from is the Northern and Central parts of the Northern Territory, including other states as well, such as S.A., W.A., and Queensland. This is the place where one and all meet, work and also make friends too. Each year different student come to do the course of teacher training here at Batchelor College. They bring their families as well to live here with them. The town of Batchelor where Batchelor College is such a nice town to see. With its lovely garden in the park (from Richards 1985:60-61).

This text immediately establishes the nature and purpose of Batchelor College (1, 2, 3), and gives the impression that this is the topic. The nature of the student body is stated (4) and then we are reminded of the head topic with a near restatement of the second sentence (5, 6). The student body theme is reintroduced and developed (7, 8, 9, 10). The passage concludes with a near mirror repetition of the first sentence (11, 12). There appears to be a form of alternation in topical progression: location (1, 2), purpose (3), student body (4), location (5, 6), student body (7), location (8), student body (9-10), location (10, 11, 12, 13).

Once again, this text follows the pattern exhibited in Texts 1, 2 and 3, although a greater syntactic complexity is evident. Discourse units 2, 5, 8, 10, 11, 12 and 13 repeat, or are synonymous, with the head topic found in Discourse Unit 1.
The final example (Text 5) is written by an Aboriginal teacher with an urban Aboriginal background undertaking post-graduate studies at Batchelor College. This student's English was developed in an ESL immersion context.

Text 5 (Adult ESL College)

1Aboriginal education is different because of the cultural differences. 2Aboriginal children do not relate very well to European teachers. 4Most are usually considered outsiders anyway. 5A lot of answers on how to teach the Aboriginal child can be found in Stephan Harris's book "Aboriginal Learning Styles." 6Stephan Harris is a man who spent a lot of time studying the learning styles and techniques of Aboriginal children at Milingimbi. 7This book which is used extensively at Batchelor College, provides a definite and positive insight on how to teach Aboriginal children. 8The failure rate at secondary schools by Aboriginal children highlights the need for Aboriginal school teachers to get the children off to a good start while at primary school (from Richards 1985:103-109).

The head topic (cultural differences in Aboriginal education) is not directly developed (1,2). The theme progresses through a number of variations including the relationship between Aboriginal children and non-Aboriginal teachers (Europeans) (3,4), reference to Harris's suggestions on how to teach Aboriginal children (5,6,7), the failure rate of Aboriginal children (8), and the need for Aboriginal children to start well (9). Note how the phrase "Aboriginal child/ren" is repeated in Discourse Units 3, 5, 6, 7, 8 and 9.

Although this text does not have exactly the same properties as the previous examples, it is clearly obvious that it does not follow the preferred rhetorical patterns of English. As mentioned above, this example was written by a student from an urban Aboriginal background who has been educated entirely in non-Aboriginal urban schools.

DISCUSSION

From the incomplete analysis of Aboriginal texts presented above, it is apparent that there is some form of discoursal patterning of Aboriginal texts which is different than what we would expect in standard English. As Figure 1 indicates, repetition and synonymy of the head topic unit are used frequently perhaps as cohesive devices which allow the speaker/writer to examine the topic from various "unrelated" viewpoints knowing that each discourse unit will be tied to the repeated head topic unit.
Figure 1
Topical Progression in Aboriginal Discourse Texts

Numbers refer to discourse units,
Dashed and dotted lines represent discourse progression through coordination or subordination,
Dashed lines represent linking due to repetition or synonymity,
R = repeat of __, S = synonymous with __, () = partly R/S __

<table>
<thead>
<tr>
<th>Text 1</th>
<th>Text 2</th>
<th>Text 3</th>
<th>Text 4</th>
<th>Text 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>:-1</td>
<td>:-2</td>
<td>:-1</td>
<td>:-1</td>
<td>1</td>
</tr>
<tr>
<td>:-3R1</td>
<td>:-4R1</td>
<td>:-R1</td>
<td>:-S1</td>
<td>2</td>
</tr>
<tr>
<td>:-5S1</td>
<td>:-6S1</td>
<td>:-5R1</td>
<td>:-6</td>
<td>3</td>
</tr>
<tr>
<td>:-7S1</td>
<td>:-7R6</td>
<td>:-6</td>
<td>:-R1</td>
<td>4</td>
</tr>
<tr>
<td>:-8S1</td>
<td>:-8R5</td>
<td>:-6</td>
<td>:-6</td>
<td>5</td>
</tr>
<tr>
<td>:-10R1</td>
<td>:-9</td>
<td>:-R1</td>
<td>:-S1</td>
<td>6</td>
</tr>
<tr>
<td>:-12R6</td>
<td>:-10</td>
<td>:-10</td>
<td>:-S1</td>
<td>7</td>
</tr>
<tr>
<td>:-13R6</td>
<td>:-11</td>
<td>:-11</td>
<td>:-1S1</td>
<td>8</td>
</tr>
<tr>
<td>:-14</td>
<td>:-12R6</td>
<td>:-12R6</td>
<td>:-12S1</td>
<td>9</td>
</tr>
<tr>
<td>:-15</td>
<td>:-13R6</td>
<td>:-13R6</td>
<td>:-13S1</td>
<td>10</td>
</tr>
<tr>
<td>:-16</td>
<td>:-17R6</td>
<td>:-17R6</td>
<td>:-17S1</td>
<td>11</td>
</tr>
<tr>
<td>:-18</td>
<td>:-19R6</td>
<td>:-19R6</td>
<td>:-19S1</td>
<td>12</td>
</tr>
<tr>
<td>:-20</td>
<td>:-21</td>
<td>:-21</td>
<td>:-2S6</td>
<td>13</td>
</tr>
<tr>
<td>:-23</td>
<td>:-24</td>
<td>:-24</td>
<td>:-5R24</td>
<td>14</td>
</tr>
<tr>
<td>:-26</td>
<td>:-27</td>
<td>:-27</td>
<td></td>
<td>15</td>
</tr>
</tbody>
</table>
Thus, in the Aboriginal English texts (3, 4, 5), what appear to native English speakers to be a list of unconnected discourse units diverging from the stated topic are actually attached to the topic by repetition and synonymity. Likewise, what appears to be a series of unnecessary repetitions are actually cohesive devices required to establish topic and tie in discourse units to that topic. These cohesive devices have been carried over from first language procedures as shown in Texts 1 and 2. Carstensen (1987) has found a similar use of repetition and synonymity in an analysis of urban Aboriginal English discourse. What are the implication of such a finding?

From a theoretical viewpoint this information contributes to the ever expanding body of data confirming the presence of L2 writing styles being influenced by L1 discourse patterns. It may be that Texts 1, 2, and perhaps 3 are part of a universal oral language narrative style. Preliminary comparisons with the oral narrative styles of Chipewyan (Scollon 1979) and Alaskan Yupik (Woodbury 1985) suggest that repetition and synonymity play a significant part in the development of topic in these two predominantly oral languages. It is interesting to note Woodbury’s comment regarding a text in English from an 8 year old Yupik boy.

CAY (Central Alaskan Yupik) rhetorical structure is clearly present, carried over to English in the form of phonological phrasing, intonation, and sentential particles. ... Such replication in English of the form and content of CAY rhetorical structure shows then just how fundamental it is to speaking in CAY communities regardless of the kind of discourse involved (Woodbury 1985:171.172).

From an educational linguistics point of view, there are a number of major implications. Krulee et al. (1979), Kintsch and Green (1978), Hinds (1987) and Eggington (1987) have examined the relationship between discoursal patterns in written texts and memory recall. It would appear that optimum memory recall occurs when the writer and reader share the same discourse framework. When there is disagreement between these two frameworks short-term memory is not affected, but long-term memory shows a significant decline. Aboriginal children, especially at a secondary level are expected to gather information through the reading of English textbooks. The evidence presented above would indicate that this is a difficult task. It is not surprising to learn that not one Aboriginal child from a predominately Aboriginal cultural background has matriculated from a Northern Territory high school.

---

As mentioned previously, Example 3 came from a collection of 33 samples of Northern Territory Aboriginal children’s writing. These samples were collected from all Northern Territory Aboriginal Community Schools operating with children from a "Stage 5" level. The samples were moderated and evaluated by a committee consisting of eight senior non-Aboriginal education officers involved in Aboriginal education. Table 1 shows the valuative labels which were attached to the texts together with the frequency of each comment. Table 2 lists a general observation on all the texts, with a selection of evaluations of individual texts.

Obviously the writing of these Aboriginal children is being evaluated using standard English rhetorical expectations. This disjunction between writer discourse pattern and reader discourse pattern has caused the readers to evaluate the texts negatively and to prescribe suggestions for further development in writing skills.

In contrast, a similar collection of children’s writing from Darwin urban schools with predominantly native English speaking students (Northern Territory Department of Education 1983), indicates that repetition, lack of structure, lack of cohesiveness and so on are not predominant features. Most rater comments focused on writing surface features (spelling, punctuation) and on content (story depth). Likewise a collection of texts written by immigrant ESL children from Europe and South-East Asia reveals expected weaknesses in proficiency, but not a tendency to repeat discourse units.

A pedagogical implication derived from the above findings revolves around a common teaching methodology used frequently throughout Aboriginal schools and possibly used throughout much of the world’s "progressive" educational systems. Often Aboriginal children are taught to write through a process entailing draft - conferencing - rewrite stages. In this model, written discourse is taught through "conferencing" where the teacher interacts with the student attempting to assist the student in discovering the appropriate rhetorical style. One wonders if conferencing can have any effect when reader (teacher) and writer (student) are operating from different rhetorical frameworks.

Forgive me for being anecdotal, but I would like to refer to an experience I had while learning the Korean language. Koreans often state the topic subject of the text once and then see no need to restate it. This subject is assumed shared knowledge and to repeat it is simply an exercise in redundancy. After being in Korea for 2 years, I believed I had fairly competent oral Korean. An occasion arose when I needed to write an important text in Korean. After producing my best effort, I gave the letter to a Korean friend who proceeded to re-write it to make it sound more "Korean". He stated politely that my text was repetitive and disorganized.
Table 1
Valuative Labels Attached to Aboriginal Texts

<table>
<thead>
<tr>
<th>Label</th>
<th>Frequency</th>
<th>Label</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>repetitive</td>
<td>3</td>
<td>disjointed</td>
<td>2</td>
</tr>
<tr>
<td>just a list</td>
<td>2</td>
<td>lack of development</td>
<td>3</td>
</tr>
<tr>
<td>no theme</td>
<td>2</td>
<td>word overuse</td>
<td>2</td>
</tr>
<tr>
<td>unrelated</td>
<td>2</td>
<td>lack of organization</td>
<td>3</td>
</tr>
<tr>
<td>lack of structure</td>
<td>3</td>
<td>lack of cohesiveness</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2
Selected Evaluations of Aboriginal Texts

General Evaluation of All Texts

[Aboriginal students] appeared to be unable to sustain the effort necessary to produce pieces of extended writing. This was obvious from the number of stories that began well but which progressively deteriorated in organization (N.T. Department of Education 1985:3).

Evaluation of Specific Texts

Even though the story is repetitive and some words have been overused, its length is such as could be expected from a child working at Stage 5 (p. 20).

Even though this is a piece of extended writing, it is repetitive, disjointed and simple words have been overused (p. 16).

This piece of writing merely lists a series of funny incidents not tied together by any main theme or story line. The writer has not understood the topic as the text of the piece is unrelated to the topic (p. 12).

Some attempt has been made to make the ideas cohesive. However, the timing is awry because of the sequencing of these ideas (p. 34).

The writer has conveyed his/her ideas and feelings, but has had some problems organizing the introduction to the story. It needs to be reread to determine who the main character is (p. 38).
However I could not see how it was repetitive, and from my point of view, his rewrite seemed rather disorganized. At that time I didn’t have the meta-linguistic knowledge to negotiate the text to an understanding of what was wrong.

Likewise, in the Aboriginal context, a student can sit down with his/her non-Aboriginal teacher and experience the same frustrations as has been recounted above. From the student’s point of view, the text is well organized - it feels good and is a reflection of the way meaning is developed in that writer’s culture.

Before proceeding with suggestions on how to teach English rhetorical styles, it may be asked if it is really necessary to require Aboriginal students to function in standard English academic registers. As mentioned previously, not one Aboriginal child from a traditional cultural background has matriculated from a Northern Territory High School. This is not surprising considering the weight of evidence suggesting that often students pass examinations not on what they know, but on how that knowledge is delivered. Indeed, Clyne (1980) indicates the ability to write in a linear, hierarchical order was a key determiner of success in Australian Higher School Certificate Matriculation examinations for non-native English speaking students.

Martin (forthcoming) has described how certain discourse styles in English are related to gaining more power in an English dominated society. He suggests that these styles are part of a "secret" language which, when mastered offers empowerment. His reference to the language being secret comes from a number of sources including Bain (1979) who quotes an Aboriginal leader’s views of educational needs:

We want them to learn. Not the kind of English you teach them in class, but your secret English. We don’t understand that English, but you do. To us you seem to say one thing and do another. That’s the English we want our children to learn (Bain, 1979).

and von Sturmer (1984) who states that:

The specific complaint, then, is that balanda (non-Aboriginal Australians) withhold the secret of their power, and that much of this 'power' is tied up with the 'big English' to which Aboriginal people are denied access. According to one interpretation, schools are failures because they fail to teach this 'power' (von Sturmer 1984:273)

Academic discourse may not be "secret" in an oral society sense, but it is a restricted and exclusive code (Kachru 1986:61), which, when mastered, leads to greater opportunity in the dominant culture.
So how can the information discussed here be helpful in assisting students towards gaining access to this code?

1. First, by making teachers aware of contrastive styles, teachers can avoid thinking in deficit terms of Aboriginal or non-standard English writing. Aboriginal students' writing is not deficient, but rather an example of applying a different set of discourse patterns to English.

2. Rhetorical patterns can actually be taught. Eggington and Ricento (1983) have outlined methods where teachers and students can come to an understanding of culturally influenced discourse patterns and then apply that knowledge to their own writing.

3. When both writer and reader, teacher and student share the same knowledge of what is expected in writing then valuable conferencing can occur and real communicative competence is achieved.

CONCLUSION

In conclusion, I would like to emphasize an indirect warning mentioned above which may be generalized to all areas of cross-linguistic and cross-cultural teaching. As teachers approach the task of assisting students from one language background towards developing meaning in a target language, there is a tendency to think in terms of a deficit model of education. It could be presumed that learners bring very little organizational ability to a text, and that all teachers have to do is feed learners the "correct" organizational patterns of the target language texts and the learners will respond with a reflection of these patterns. Hopefully, this paper has shown that students bring a whole range of first language influenced preferences for developing meaning to the text. In addition, these preferences greatly influence the development of meaning in the second language. The suggestions given above for dealing with this situation are but a few which could be developed once teachers become aware of the cultural depth of each student.

This paper has focused on the English of Aboriginal people. Many of the above concepts can be generalized to most cross-linguistic contexts. Since it would be a safe assumption to conclude that most ESL students have a discourse organizational style not in harmony with those patterns preferred in English, the conclusions reached in this paper have a certain validity in most ESL teaching situations.
REFERENCES


____ (1983) NT Children's writing: years 5 and 7 moderated samples. Darwin: N.T. Department of Education.


Shimpo, M. (1985) Communication processes between the


Here are two riddles:

"A man is walking down the street one day when he suddenly recognizes an old friend whom he had not seen in years walking in his direction with a little girl. They greet each other warmly and the friend says, 'I married since I last saw you, to someone you never met, and this is my daughter, Ellen.' The man says to Ellen, 'You look just like your mother.' How did he know that?

"A boy and his father were driving when suddenly a large truck careened around a corner and hit their car head-on. The car was crushed, and when their bodies were removed from the wreck the father was already dead. The son, badly injured but still alive, was rushed to the hospital, where hasty preparations were made for immediate surgery. As the boy was brought in for the operation, the surgeon saw him and said, 'I can't operate, that's my son.' How is that possible?" (Moulton, 1981, pp. 100-101).

Sexist language has only recently become an issue. It wasn't until the mid Seventies that people, mostly women, gathered together to protest against sexist language.

Else and Sanford (1987) divided the term "sexist language" into two categories. The first category is the selecting of content, which assigns roles or characteristics on the basis of sex in ways that create and reinforce negative images of women, sex-role stereotypes, or unequal treatment of the sexes. The other category is the use of language forms and styles that excluded women, distort information about human behavior or limit perceptions of behavioral options by using masculine terms as if they were inclusive," (p. 52).

Have you solved the riddles yet? If not, carefully read them again. One would tend to think that in the first riddle the man's friend was a male and that the surgeon was also male.

The pronouns "he", "his", and "him" have always referred to a male. However, according to the usage authorities, they can also
refer to either a male or a female when the sex of the subject is unknown or applying to both sexes.

In 1850, Parliament wanted to shorten the language used in its documents. Thus, it announced that

in all acts words importing the masculine gender shall be deemed and taken to include females, and the singular to include the plural, and the plural the singular, unless the contrary as to gender and number is expressly provided (Evans and Evans, 1957, p. 221).

Styles and attitudes are changing, especially now in the late 1980's. Society is altering its view about many things such as marriage, family, and children. Can we include in these changes the mandate to avoid sexist language? Just how many of the people are against "he" as a gender neutral pronoun? Do they always think of a male when "he" is used generically? If they do try to avoid using "he", what do they use?

The research question that is asked is "In writing, is "he" considered a gender neutral pronoun among university students?" Writing was chosen as the focus of research rather than speaking because the use of "he" may be more controversial in writing. University students were chosen as participants because they are immersed in the generic pronoun problem every time they write a paper. "He", "his", and "him" will be used interchangeably along with "they", "theirs", and "their" and "he or she", "his or hers", and "him or her" because they are all pronouns.

By studying the use of "he" as a gender neutral pronoun, it is anticipated that the results can be used to measure the influence of the sexist pronoun controversy, and act as a guide to those who are still confused about which singular pronoun to use.

In order to gain a tighter grasp on the subject, let's review what usage authorities say about the use of "he" as a generic pronoun, and also review the studies that have been conducted dealing with "he".

The purpose of usage authorities is to describe the use of language in society. Most deal directly with controversial subjects such as this.

1) First of all, The American Heritage Dictionary (1975), considered conservative, was compiled in reaction to Webster's Third International Dictionary which was very liberal. The AHD summarizes the neutral use of "he" as representing any person whose sex is not specified (Morris, 1975, p. 606).

2) Copperud's American Usage: The Consensus (1970) summarizes the judgments of eight usage dictionaries and seven general dictionaries on various points of usage. It is stated that "it
is a well-established convention that the masculine form alone is taken as applying to both sexes," (pp. 124-125).

3) Ebbitt and Ebbitt (1982), both liberal and conservative, say that "traditionally, the masculine pronoun "he" is used with indefinite pronouns like "anyone" and "everybody" and with noun antecedents that may refer to either men or women. But feminists find this usage a prime example of sexist language," (pp. 460-461).

4) Evans and Evans (1957), a brother and sister team who are rather liberal and make extended explanations of their opinion, assert that "he" can be used for either men or women (p. 221).

5) Pollett (1966) and Bernstein (1966), who are conservatives; Bryant (1962), who extracted most of her material from original writing research; and Fowler (1927 and 1965), a British usage authority, do not address the problem of using "he" as a generic form.

6) Morris and Morris (1975), who polled a panel of conservatives for usage judgments, state that "there have been rising protests . . . against the continuation of the traditional use of 'him' and 'his' when the person referred to is unidentified as to sex." They proceed to offer some suggestions, and later resolve that the problem is unsolved (p. 298).

7) Webster's New Collegiate Dictionary (1973) says that "he" can be "used in a generic sense of when the sex of the person is unspecified," (p. 527).

8) Webster's New World Dictionary (1970) says that one of the definitions of "he" is "the person . . . the one, anyone," (p. 643).

9) Finally, Funk & Wagnall's Standard College Dictionary (1973) also says that "he" can be "that person, anyone, one," (p. 617).

Now, we know what the usage authorities say about the use of "he" when a person is unidentified. They say that "he" can be used as a generic pronoun. Some also recognize that there is a problem in society about using "he" in that way. Let's look at the studies that have been conducted about the use of "he", "he/she", and "they" as gender neutral pronouns.

The following is a compilation of studies which have been conducted within the last decade and inquire about the use of "he".

1) As cited in Else and Sanford (1987), Schneider and Hacker (1973) conducted a study using college students. The participants were asked to read a text and choose pictures illustrating the text. One group of subjects were given "he"/"man" terms and the other, gender-free labels (which were
2) Martyna (1978) conducted a study using twenty males and twenty females. The subjects were given forty-eight sentence fragments. Eighteen were critical elements and thirty, filler. Half of them were given orally and the other half, written. Of the eighteen critical fragments, six were male-related, six were neutral, and six were female-related. The following are examples of the sentence fragments used:

**Male:** Before a judge can give a final ruling, ____ must weigh the evidence.

**Neutral:** When a person loses money, ____ has got to feel bad.

**Female:** After a nurse has completed training, ____ goes to work.

The results of Martyna's study indicated that the choice of pronoun was influenced by the presumed sex of the subject regardless if the sentence was given orally or written nor if the participant was male or female. Martyna concludes that "'he' is an ambiguous term which often allows a specifically male interpretation to be drawn from an intended generic usage," (pp. 131-138).

3) Hyde (1984) conducted two experiments to determine age differences between children and their responses to sexist language. The first experiment was taken of first, third, and fifth graders, and college students. They were given four tests: 1) They had to make up a story about a person. A third of the subjects were given pronouns of "he", a third were given "they", and the last third were given "he or she". 2) They had to fill-in sentences with pronouns. 3) Sentences were read to them with pronouns and they had to say if the sentences were correct or incorrect. 4) They were asked about the gender-neutral rule of "he". Because of the results of the first experiment, Hyde concluded that "he" is not gender-neutral. Twelve percent of those given the pronoun "he" said it included females; eighteen percent of those the pronoun "they" said it included females, and forty-two percent of those given the pronouns "he or she" said it included females. Twenty-eight percent of the children knew the gender-neutral rule and eighty-four percent of the college students knew the rule.

Experiment Two was almost the same as Experiment One except that only third and fifth graders were used, the pronoun "she" was added to the story in Test One, and given an occupation, the subjects were asked to rate how well a female could perform the job on a three-point scale. In Test One (the story), when "she"
was used, seventy-seven percent said the pronoun included females. In the test about job ratings, those given the pronoun "he" gave women a low score on accomplishment, those given the pronoun "they" gave women an intermediate score on accomplishment, and those given "she" gave women a high score on accomplishment.

As a result of these experiments, Hyde asked the question: "Does sexist language produce sexist thought or does sexist thought produce sexist language?" She concluded that sexist thought produces sexist language because even when they used the pronoun "they", the children still thought of males (pp. 697-706).

4) Moulton, Robinson, and Elias (1978) researched the understanding of "he" by giving a statement to college students and having them write about it. The participants were 226 males and 264 females which were randomly assigned to six groups. They were asked to make up a story about a fictional person who fits the statement's theme. The participants were asked not to write about themselves. Three groups were given the following statement with a different pronoun ("his", "their", or "his or her") given to each group: "In a large coeducationally institution the average student will feel isolated in introductory courses." Each of the other three groups also had different pronouns and the following statement: "Each person knows when appearance is unattractive." The results were that when "his" was used, thirty-five percent wrote about females; "their", forty-six percent; and "his or her", fifty-six percent. The researchers concluded that "he" failed to be gender neutral (pp. 1032-1036).

5) Miller and Swift (1980) say that the studies confirm that in spoken usage--from the speech of young children to the conversation of university professors--"he" is rarely intended or understood to include "she". On the contrary, at all levels of education, people whose native tongue is English seem to know that "he", "him", and "his" are gender-specific and cannot do the double duty asked of them (p. 37).

As we have seen so far, Miller and Swift's opinion of "he" as a gender-specific pronoun and not a gender-neutral one remains constant. However, they do not state which studies confirm their hypothesis.

6) This final article by Cole and Hill's (1983) was the only study found to support "he" as a gender neutral pronoun. In their study entitled, "Do Masculine Pronouns Used Generically Lead to Thoughts of Men?", they conducted five experiments with a varied number of participants in each experiment: sixteen to fifty-two men and sixteen to fifty-one women. In each experiment, the participants were given job descriptions for
certain occupations such as recreation worker and secondary school teacher using "he", "they", "he/she" alternately. The first experiment used gender-oriented jobs such as mechanical engineer and flight attendant to validate semantic differences. In the following experiment, each participant was given a booklet which used only one pronoun consistently. Cole and Hill found no evidence that masculine pronouns brought thoughts of men; however, some evidence was found when "he" was used with "man". The use of the equal pronoun ("they") did not increase the subjects' tendency to think of women (p. 737-750).

Again, the research question asks "In writing, is "he" considered a generic pronoun among university students? The participants were students attending Brigham Young University in Provo, Utah, during the 1988 winter semester. The students were not randomly selected. The information was taken from students in dormitories, classes, apartment complexes, and a park.

Thirty males and thirty females were asked to participate in the study. There were twelve freshmen, fourteen sophomores, seventeen juniors, fourteen seniors, and three graduate students. There were nineteen students in the College of Family, Home, and Social Sciences; seven undeclared majors; four students in the College of Physical and Mathematical Sciences; five in the College of Business; eight in the College of Engineering and Technology; seven in the College of Humanities, five in the College of Biology and Agriculture; one in the College of Physical Education; three in the College of Education; and six in the College of Fine Arts and Communications. The ages ranged from seventeen to thirty with the average being 21.7 years.

The research was conducted in survey form because it can be self-administered; is easier to reach a wide population; does not take as much time as a personal interview; and if well constructed, has no interviewer effect.

Moulton, Robinson, and Elias's (1978) data collection was used as an example for this study. This survey had two parts. In Part I of the survey, the participants were asked to provide personal information such as age, sex, year in school, and major. The majors were then categorized into the various colleges. The participants were then instructed to write a paragraph about a fictional character according to the following theme: "At Brigham Young University, a new student can feel alone during (his, his or her, their) first semester." One-third of the forms had "his first semester"; one-third, "his or her first semester", and one-third, "their first semester".

After the participants finished Part I, they were each given Part II and were asked to answer "yes" or "no" to the following question: "Did you know that "he" could be used as a generic pronoun referring to both male and female?"
After conducting the survey, a chi-square test of significance was used to determine the chances of the results being correct, or whether or not they were significant. A ninety-five percent confidence level was used for the test. There were varying degrees of freedom and varying chi-square values which were needed for the hypothesis to be significant.

The main questions about the survey were . . .

1) Did the writing of male or female depend on the given of "his", "his/her", and "their"?

Four supplementary questions about the survey were . . .

2) Did the knowledge of "he" as a gender neutral pronoun depend on the sex?

3) Did the knowledge of "he" as a gender neutral pronoun depend on year in school?

4) Did the knowledge of "he" as a gender neutral pronoun depend on major?

5) Did the writing of male or female depend on the sex?

QUESTION ONE: Did the writing of male or female depend on the given of "his", "his/her", and "their"?

The results of Question One are summarized in Table One:

<table>
<thead>
<tr>
<th>He</th>
<th>She</th>
</tr>
</thead>
<tbody>
<tr>
<td>His</td>
<td>15/75%</td>
</tr>
<tr>
<td>Their</td>
<td>13/65%</td>
</tr>
<tr>
<td>His or Her</td>
<td>11/55%</td>
</tr>
</tbody>
</table>

According to Table One, when given "he", seventy-five percent wrote about a male and twenty-five percent wrote about a female. When given "their", sixty-five percent wrote about a male and thirty-five percent wrote about a female. When given "his or her", fifty-five percent wrote about a male and forty-five percent wrote about a female. It seems that the pronouns were not equally distributed among the categories. In addition, "he" was more often thought of as a male than "their" or "his or her". It also can be concluded that when "his or her" was given, people thought almost equally about males and females. Maybe this can reflect the influence of the sexist pronoun controversy.

In the statistical analysis of Question One, a ninety-five percent confidence level and two degrees of freedom were used. Because the chi-square value needed to be greater than 6.0 and the actual value was 1.758, no conclusion could be made. It
could not be statistically determined that the writing of male or female depended on the given of "his", "his or her", and "their".

Because this survey was modeled after the one in Moulton, Robinson, and Elias (1978), it was speculated that maybe the results of this survey could be significantly similar to the ones of Moulton et al. Therefore, a large sample test of a population proportion was conducted. Each of the percentages of "his", "their", and "his or her" of this survey were compared with those of Moulton et al.'s survey. The hypothesis would be rejected if the z score was greater than 1.96 or less than -1.96. The z score of "his" was -0.937. The z scores of "his or her" and "their" were -1.03. Therefore, the hypothesis was not proved incorrect. Maybe we can assume to be true that the results of this survey were similar to the ones of Moulton, et al.

QUESTION TWO: Did the knowledge of "he" as a gender neutral pronoun depend on the sex?

The results of Question Two are summarized in Table Two:

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>28/93%</td>
<td>2/6%</td>
</tr>
<tr>
<td>Male</td>
<td>21/70%</td>
<td>9/30%</td>
</tr>
</tbody>
</table>

According to Table Two, ninety-three percent of the females said "yes" that they understood that "he" could be used as a generic pronoun and six percent of the females said "no". Seventy percent of the males said "yes" and thirty percent said "no". It seems that more females had a understanding that "he" could be used as a generic pronoun.

In the statistical analysis of Question Two, a ninety-five percent confidence level and one degree of freedom were used. The chi-square value needed to be greater than 3.84 and the outcome was 5.454. Thus, we can be ninety-five percent confident that knowledge of "he" as a generic pronoun depended on the sex.

QUESTION THREE: Did the knowledge of "he" as a gender neutral pronoun depend on year in school?

The results of Question Three are summarized in Table Three:

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshman</td>
<td>8/73%</td>
<td>3/27%</td>
</tr>
<tr>
<td>Sophomore</td>
<td>12/80%</td>
<td>3/20%</td>
</tr>
<tr>
<td>Junior</td>
<td>12/80%</td>
<td>3/20%</td>
</tr>
<tr>
<td>Senior</td>
<td>14/88%</td>
<td>2/12%</td>
</tr>
<tr>
<td>Graduate</td>
<td>3/100%</td>
<td>0/00%</td>
</tr>
</tbody>
</table>
According to Table Three, seventy-three percent of the freshmen said "yes" that they understood that "he" could be used as a generic pronoun and twenty-seven percent said "no". Eighty percent of the sophomores said "yes" and twenty percent said "no". Eighty percent of the juniors said "yes" and twenty percent said "no". Eighty-eight percent of the seniors said "yes" and twelve percent said "no". One hundred percent of the graduates said "yes". It seems that the higher the year in school, the more likely one understood that "he" could be used as a generic pronoun.

In the statistical analysis of Question Three, a ninety-five percent confidence level and four degrees of freedom were used. Because the chi-square value needed to be greater than 9.50 and the actual value was 1.672, no conclusion could be made that the understanding of "he" as a generic pronoun depended on year in school.

QUESTION FOUR: Did the knowledge of "he" as gender neutral pronoun depend on major?

The results of Question Four are summarized in Table Four:

<table>
<thead>
<tr>
<th>Major</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering and Technology</td>
<td>6/60%</td>
<td>2/40%</td>
</tr>
<tr>
<td>Business</td>
<td>4/67%</td>
<td>2/33%</td>
</tr>
<tr>
<td>Fine Arts and Communications</td>
<td>6/75%</td>
<td>2/25%</td>
</tr>
<tr>
<td>Biology and Agriculture</td>
<td>3/75%</td>
<td>1/25%</td>
</tr>
<tr>
<td>Physical and Mathematical Sciences</td>
<td>4/80%</td>
<td>1/20%</td>
</tr>
<tr>
<td>Family, Home, and Social Sciences</td>
<td>12/86%</td>
<td>2/14%</td>
</tr>
<tr>
<td>Undeclared</td>
<td>6/86%</td>
<td>1/14%</td>
</tr>
<tr>
<td>Law School</td>
<td>1/100%</td>
<td>0/00%</td>
</tr>
<tr>
<td>Physical Education</td>
<td>1/100%</td>
<td>0/00%</td>
</tr>
<tr>
<td>Education</td>
<td>3/100%</td>
<td>0/00%</td>
</tr>
<tr>
<td>Humanities</td>
<td>6/100%</td>
<td>0/00%</td>
</tr>
</tbody>
</table>

According to Table Four, sixty percent of those in engineering and technology said "yes" that they understood that "he" could be used as a generic pronoun and forty percent said "no". Sixty-seven percent of those in business said "yes" and thirty-three percent said "no". Seventy-five percent of those in fine arts and communications and those in biology and agriculture said "yes" and twenty-five percent said "no". Eighty percent of those in physical and mathematical sciences said "yes" and twenty percent said "no". Eighty-six percent of those in family, home, and social sciences and those with no declared majors said "yes" and fourteen percent said "no". One hundred percent of those in law school, physical education, education, and humanities said "yes". It seems that those majors in education and humanities were more aware that "he" could be used as a generic pronoun.
In the statistical analysis of Question Four, a ninety-five percent confidence level and ten degrees of freedom were used. Because the chi-square value needed to be greater than 18.30 and the actual value was 5.509, no conclusion could be made that knowledge of "he" as a generic pronoun depended on major.

QUESTION FIVE: Did the writing of male or female depend on the sex?

The results of Question Five are summarized in Table Five:

<table>
<thead>
<tr>
<th></th>
<th>He</th>
<th>She</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>27/90%</td>
<td>3/10%</td>
</tr>
<tr>
<td>Female</td>
<td>12/40%</td>
<td>18/60%</td>
</tr>
</tbody>
</table>

According to Table Five, ninety percent of the males wrote "he" and ten percent wrote "she". Forty percent of the females wrote "he" and sixty percent wrote "she". It seems that the writing of male or female depended on sex. It also seems that more females were inclined to write about a male than males were to write about a female.

In the statistical analysis of Question Five, a ninety-five percent confidence level and one degree of freedom were used. The chi-square value needed to be greater than 3.84 and the outcome value was 16.484. Thus, we can be ninety-five percent confident that the writing of male or female depended on the sex.

The following is a summary of the survey results of the five questions:

1) Did the writing of male or female depend on the given of "his", "his/her", and "their"? No, it could not be statistically determined that the writing of male or female depended on the given of "his", "his or her", and "their". However, we may be able to assume that the survey results are statistically similar to those of Moulton, Robinson, and Elias (1978).

2) Did the knowledge of "he" as a gender neutral pronoun depend on the sex? Yes, we can be ninety-five percent confident that knowledge of "he" as a generic pronoun depended on the sex.

3) Did the knowledge of "he" as a gender neutral pronoun depend on year in school? No, no conclusion could be made that the understanding of "he" as a generic pronoun depended on year in school.

4) Did the knowledge of "he" as a generic pronoun depend on
major? No, no conclusion could be made that knowledge of "he" as a generic pronoun depended on major.

5) Did the writing of male or female depend on the sex? Yes, we can be ninety-five percent confident that the writing of male or female depended on the sex.

The outcome of the statistics for Question One were surprising because almost all of the other research conducted and summarized in the review of literature supported that "he" is thought of as a generic pronoun. It was expected that the survey results from Question One would be significant. However, it is believed that if a random sample was taken and a larger sample of each category (such as thirty males and thirty females for each "his", "their", and "his or her") the results may have been significant.

The results of this survey should not be considered definite because no random sample was taken and the categories were too small. In spite of this, because of the large sample test of population proportion, we can assume that the survey results support those of Moulton, et al. (1978), and that "he" is most often thought of including only males and is not a generic pronoun.

Now, for those people who feel they may a problem with using "he" as a generic pronoun, many of the cited studies offer suggestions of how to avoid using "he":

1) First of all, Miller and Swift (1980), Frank and Anshen (1983), Morris and Morris (1975), Ebbitt and Ebbitt (1982), and Shear (1981) suggest using "he or she". However, Copperud (1970) and Archibald (1970) believe that "he or she" is too clumsy.

2) Miller and Swift (1980) and Ebbitt and Ebbitt (1982) say that "he/she" or "s/he" would be all right too.

3) Also, Miller and Swift (1980), Frank and Anshen (1983), Ebbitt and Ebbitt (1982), Copperud (1970), Archibald (1978), Evans and Evans (1957), and Bernstein (1977) advocate the use of the plural pronouns "they" or "their" even after a singular antecedent such as "everyone", "everybody", "someone", and "somebody". There is one opponent to this usage: Copperud (1979) who himself contradicts his recommendation of "they" in 1970.

4) Miller and Swift (1980) and Frank and Anshen (1983) suggest using "she" as the generic pronoun.

5) Another suggestion was made by Morris and Morris (1975) and Shear (1981) to use "one" instead of "he".

6) Pluralizing is a good way of avoiding the neutral pronoun. Then no one realizes that writers have a problem with sex fairness (Miller and Swift, 1980; Frank and Anshen, 1983; Shear, 1981; and Nielsen, 1984).
7) Some authors as cited in Frank and Anshen (1983), and Miller and Swift (1980), have introduced new pronouns which are nonsexist. In the 1800's, Charles Converse proposed the use of "thon" (that one). It was in the dictionaries until the 1950's. Other such proposed pronouns include "co", "E", "tey", and "hesh". June Arnold used "na" in The Cook and the Carpenter. Marge Piercy used "pa" for a short form of person in Woman on the Edge of Time. Ray A. Killian, author of Managers Must Lead! used "hir".

8) Finally writers must above all be consistent while using the pronoun forms (Nielsen, 1984; Shear, 1981).
References


BRIGHAM YOUNG UNIVERSITY SURVEY
PART I

PERSONAL INFORMATION:
Age____
Sex____
Year in school_______________
Major_______________________

INSTRUCTIONS: Write a paragraph of a fictional character according to the following theme. PLEASE DO NOT WRITE ABOUT YOURSELF.

THEME: At Brigham Young University, a new student can feel alone during his or her first semester.
PERSONAL INFORMATION:

Age

Sex

Year in school

Major

INSTRUCTIONS: Write a paragraph of a fictional character according to the following theme. PLEASE DO NOT WRITE ABOUT YOURSELF.

THEME: At Brigham Young University, a new student can feel alone during his first semester.
BRIGHAM YOUNG UNIVERSITY SURVEY
PART I

PERSONAL INFORMATION:
Age____
Sex____
Year in school__________________
Major______________________

INSTRUCTIONS: Write a paragraph of a fictional character according to the following theme. PLEASE DO NOT WRITE ABOUT YOURSELF.

THEME: At Brigham Young University, a new student can feel alone during their first semester.
ANSWER: Did you know that "he" could be used as a generic pronoun referring to both male and female?
SUPERMARKET LEXICOGRAPHY
OR
FINDING NEW TERMS TO PUT IN THE DICTIONARY

Robert L. Good
ALPNET

For the past five years or so I have been the chief lexicographer at a local company called ALPNET that produces computer-assisted translation software as well as other natural language processing software. Part of the work I have been doing has involved the creation of monolingual English linguistic databases and bilingual dictionaries to support our software applications.

Creating dictionaries from scratch is not an easy task. The first obvious question that arises is "What am I going to put into this dictionary?" The answer, of course, depends on a multitude of other questions: What is the dictionary going to be used for? Who will use the dictionary? What features will allow the users easy access to the contents of the dictionary? What medium will it end up in? What are the size limitations? What sources should be consulted, and how, in order to create the dictionary? What format will the physical entries take? Will outside experts be required? What is the budget for this dictionary? What are the deadlines involved and how will they affect everything else?

These are questions every editor must answer to produce the best dictionary possible given the constraints as represented by the answers to these questions. If the dictionary is a wholly new creation, then identifying the list of headwords to be treated is a significant undertaking. Even if one is just making a revision of a pre-existing dictionary, it is still not trivial. It is likely that it will be necessary to add new terms to the dictionary to increase the desired coverage. Where do these lists of terms for consideration come from?

Traditionally, publishers of dictionaries have access to individuals whose job or hobby it is to read through copious amounts of text—newspapers, magazines, books of all sorts, professional journals, etc.—looking for examples of new terms or old terms used in novel ways. With the advent of massive amounts of electronic corpora and corpus analysis tools for analyzing them, it is now possible to have the computer do much of the work, at least in the step of identifying new terms. Much of my work at ALPNET has
centered around the design of computer-based corpus analysis tools for identifying new terms, both single-word and multiple-word.

The multiple-word terms have been of particular interest because, when translating from one language to another, word-for-word translations are rarely ever adequate. At ALPNET we have explored syntax parsing algorithms and implemented a system for identifying multiple-word terms. This has made it possible to identify potentially useful strings of words for automatic inclusion into our translation dictionaries.

But, in addition to manually reviewing texts or analyzing them using a computer, there is perhaps another source of new terms for the dictionary that does not appear to have been tapped yet. And that is the local supermarket. Most people probably are not aware that they cannot buy toilet paper, soda crackers, lunch meat, or dishwashing soap at any of their local supermarkets. They might get what they think is toilet paper, but this is not the label that appears on the packaging. Of course, once we are told that the advertiser's name for the product is not 'toilet paper' we instinctively go through some mental gymnastics to conjure up a suitable euphemism. The one that is used is 'bathroom tissue'.

Similarly, 'lunch meat' never appears on the package; the specific meat is always indicated. 'Dishwashing soap' is really 'dishwashing liquid' and 'soda crackers' are typically 'saltine crackers', at least as long as they have salt on them. One manufacturer called its regular and low salt crackers 'saltine crackers', but its unsalted crackers were merely 'crackers'.

None of these products claims to be toilet paper, lunch meat, dishwashing soap, or soda crackers. What we see here is a discrepancy between at least one common name and the generic name manufacturers have chosen to label these products. We have all seen this phenomenon with brand names that have become the common name for the product, like linoleum. We have also seen other cases where many of us use brand names as if they were generic names, even though the brand name is not yet in the public domain. Examples include Kleenex, Band-Aids, and Xerox machines, which are more generically referred to as facial tissues, bandages, and photocopying machines. These generic names are actually quite descriptive. But this is not always the case. Occasionally a manufacturer will pick a name that is really not descriptive at all. For example, if someone sent you to the store to buy some 'plastic strips' what would you bring back?
The term 'plastic strip' happens to be Curity's name for what I always call 'band-aids'—regardless of the brand—and what Johnson & Johnson insists on calling 'plastic bandages' so that they can retain a proprietary right to the name 'Band-Aid'. Curity's name is far from transparent.

Dictionaries seem to do a more thorough job of identifying terms that are proprietary but may also have a life of their own as generic terms in popular usage. This attention to detail is not so evident for non-proprietary names like 'bathroom tissue'. In fact, there appears to be quite a gap between the general monolingual desk dictionary's content with regard to these terms and the products on the shelves at the supermarket.

Appendix B lists the products and terms I checked the shelves for at a large local supermarket and then looked for in the dictionaries listed in Appendix A. I was sometimes surprised by the terms underrepresented in the dictionaries. Terms like 'eye drops', 'light' (meaning low in calories), 'ice-cream sandwich', 'lunch meat', 'Q-Tips', and several others seemed to be worthwhile terms.

On the other hand I was somewhat surprised at some of the generic terms chosen by manufacturers to call their products. Everyone knows what an ice-cream cone is, even though none of the dictionaries explicitly state that it does not have to be in the shape of a cone. I call the flat-bottomed ones cones, as I do the truly cone-shaped sugar cones. But the manufacturers of the flat-bottomed ones prefer the term 'cup' to describe their product. That seemed a little strange to me.

After looking at the various products in Appendix B and their commercial names, what are we to conclude? Should any of these terms appear in a dictionary of English? This is a difficult question. The mere existence of a term does not mean that it ought to be included in a dictionary. And if the term is composed of more than one word, it is even more difficult to decide if it warrants individual attention. The size, scope, and budget for the dictionary naturally restrict the number of terms that can be included and the depth of their treatment. The frequency and distribution and durability of each term must also be considered. Nonce words do not make good entries in a general dictionary. Neither do terms so specialized that no one is likely to ever come across them. But a good selling point for a dictionary can be the number of "new terms" that have been added since the last edition.

Simon and Schuster has recently published the third college edition of their New World Dictionary. The dust jacket mentions that 5,000 new terms have been included. I did not notice in the introductory material any discussion of these
terms, where they came from, or by what criteria they were admitted into the printed lexicon, but there are 5,000 of them and some of them are probably ones whose exclusion from earlier editions has puzzled us for years. The factors mentioned before concerning frequency, distribution, and durability, as well as the publishing requirements no doubt influenced the decision to include or exclude each term. The publishers very likely had many more new terms than they had room for or thought it wise to include without additional citations.

Some of the terms that I found browsing the shelves at the supermarket may not seem very important, but they have great exposure and distribution. Some of them are possibly the kinds of terms one would like to be able to look up in a dictionary. For example, when speaking of ground beef, we know that there are at least three kinds, differentiated by fat content: regular, lean, and extra lean. I don’t know if there is a standard throughout the beef industry, but at one large local supermarket it states right on the package what the upper limits for the percentages are. I thought it was interesting that not only are the grades of hamburger missing from the six dictionaries I consulted, but also the word ‘ground beef’.

Another specific set of terms that might be included are the ones describing what I call sliced cheese products. These are the individually wrapped "cheese" slices, some of which the supermarkets do not even refrigerate, which makes one wonder what is in them. The range of names for these are varied and some are so ominous sounding that I would really like to know what the differences are. Some of the names of these products include: pasteurized process (American/Swiss) cheese; pasteurized process cheese product; pasteurized process cheese spread; pasteurized process cheese food; pasteurized process cheese food substitute. The first one sounds like real cheese. I have no idea what I should think about the last one. Perhaps these names are dictated by law because of the content. If so, are they sufficiently standardized that they could be treated in a dictionary?

Dictionaries with more precise descriptions of consumer products could provide a useful service to the average consumer. Such descriptions might specify the different grades of ground beef, the general contents of cheese products, and percentages of butter fat in ice cream, ice milk, and light ice milk, etc.

I recognize that the editorial process requires that some terms be excluded in preference to others. The desk dictionaries that I consulted were certainly not unabridged dictionaries and there is only so much time, space, and money that can be devoted to making a dictionary. Even unabridged dictionaries have their limitations. I am not
claiming that any of these particular terms demands treatment to the exclusion of some other term. I am merely recommending one more place, and a very common one at that, to look for possible new words to put into the dictionary.
Appendix A

Dictionaries Consulted for this Study

Desk dictionaries


Webster’s Ninth New Collegiate Dictionary. Merriam-Webster, 1984. 160,000 entries and 200,000 definitions.

Unabridged dictionaries


Webster’s Third New International Dictionary, Unabridged. Merriam-Webster, 1981. 460,000 entries.

Appendix B

Supermarket Product Names and their Distribution Among the Six Reference Dictionaries

Parentheses indicate possible generic names for the products that are evidently not used on packaging or they contain explanatory notes. The terms marked "general" often appeared in the dictionary with only their more general definitions rather than with the word sense desired for the specific context given here.

<table>
<thead>
<tr>
<th>Term</th>
<th>Number of Dictionaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band-Aids (Trademark)</td>
<td>6</td>
</tr>
<tr>
<td>band-aids</td>
<td>2</td>
</tr>
<tr>
<td>bandages</td>
<td>6</td>
</tr>
<tr>
<td>plastic strips</td>
<td>0</td>
</tr>
<tr>
<td>pasteurized process (American/Swiss) cheese</td>
<td>0</td>
</tr>
</tbody>
</table>
pasteurized process cheese product 0
pasteurized process cheese spread 0
cheese spread 1
pasteurized process cheese food 0
clean food 1
pasteurized process cheese food substitute 0
cotton balls 0
cosmetic puffs 0
puffs 5
dishwashing liquid 0
dish detergent 0
dishwashing detergent 0
liquid detergent 0
eye drops 2
eyedropper 6
ground beef 0
  regular 0
general
  lean 0
general
  extra lean 0
general
hot dog 6
frank 6
wiener 6
ice cream 6
ice milk 6
light ice milk 0
light 2
general
sorbet 5
ice cream sandwich 0
(ice-cream cone) 5
cups 0
general
ice cream cups (flat bottom cones) 0
sugar cones (pointed cones) 0
cone 3
general
facial tissues 2
tissues 6
imitation krab flakes 0
imitation krab salad 0
krab 0
(lunch meat) 1
cold cuts 6
luncheon meat 3
paper towels 0
towels 6
cotton swabs 0
Q-Tips 2
swabs 6
soda cracker 6
saltine cracker 0
cracker 6
saltines 6
bathroom tissue 1
bath tissue 1
(toilet paper) 6
(toilet tissue) 3
Various stages of language learning and various language tasks suggest different learning tools. This paper does not address tools for the beginning student (i.e. first year). Rather, it describes a tool for the intermediate or advanced student who has mastered a core vocabulary and basic grammar and who wants to improve in second language writing.

There are at least two schools of thought on second language writing. One school claims that the student should stay entirely within active vocabulary. This nearly eliminates the need for a bilingual dictionary, since any word which is not already known in the second language will not be used. Vocabulary building tasks place in structured vocabulary exercises, not in writing. Another school suggests that the intermediate or advanced student should stretch while writing. This means that if an expression is not known in the second language, the student should find it using a native language expression and bilingual dictionary. It seems obvious that in this approach, the time in the native language should be minimized. This suggests that the less time spent looking up an expression the better.

WHAT IS AVAILABLE NOW

For the second approach, especially if the student is already using a word processor while writing, an electronic dictionary seems a natural alternative to a paper dictionary. An electronic dictionary product called Collins On-line(tm) will be described and demonstrated. It will be available in May 1989 in the U.S.A. This product solves the problem of expanding a demonstration dictionary into a full-scale dictionary, by incorporating an electronic version of a Collins(tm) bilingual dictionary. One advantage of the Collins dictionaries over those of other publishers is the careful Collins "signposting" which helps the student select the appropriate translation equivalent.

ENHANCEMENTS NEEDED

Future versions of electronic dictionary software should permit easier search of expressions, especially when the headword of the expression is not known. Approximate spelling, base form reduction, and grammar helps would also be useful. Although the current version uses a simple character-based interface on an IBM-PC type machine (to conserve resources and be available on the largest possible installed hardware base) it would be clearly desirable to develop a graphics-based interface as well.
Recognizing the Need

For years poets, writers, linguists, scrabble players, crossword puzzle addicts, language materials developers, and other language lovers have had frequent occasion to search for words having particular sounds or sound combinations. The poet may search for a word beginning with the right consonant to complement an alliterative line, or a word with particular vowels to complete the assonance in a line, or perhaps he may search for the right final syllable for a perfect or identical rhyme. The language materials developer may need to collect examples of particular consonant clusters or vowel combinations to help foreign students practice English phonological patterns. The linguist may be interested in statistical data such as frequency of occurrence of particular sounds, or occurrence of particular phonotactic patterns, for which English orthography is inadequate. Even "Dear Abby" needed such information to respond to a writer who asked whether there were more than two words (angry and hungry) that ended in the sounds "gry." Abby could not think of others, but offered an anecdote about a lady who approached George Bernard Shaw with the observation that there was only one word in English beginning with the letters su that had the sound [sU]. That word was sugar. George Bernard Shaw paused for a moment and replied, "Are you sure?"

Each of these circumstances illustrates the need for a way to retrieve words from the English lexicon in a way more rapid than introspection. To meet this need, rhyming dictionaries have been published, but they are generally limited to word-final rhymes, and are relatively time-consuming. Of course the ideal solution would be a computer program that would enable a user to select any sound(s) in any word position(s), and let modern technology retrieve the words with the corresponding sounds, be they vowels for end rhyme or assonance, or consonants for alliteration or consonance. Such a program should at least allow for the use of key words as models, the selected sounds of which would be matched by the computer in selecting corresponding words. Ideally, it should also allow the user to select sounds independent of any word, and thereby retrieve words containing those sounds in the phonotactic positions specified by the user. Finally, it would be a bonus to have such a program interfaced with a thesaurus, so that one may search for a word within a given semantic range containing a
particular sound or sound combination. Such a feature would help those afflicted with the "tip of the tongue" phenomenon.

**Developing a Solution: Data Base**

These are the goals of the current project which we are presenting today. In order to accomplish such a task, two things must be in place: First of all, there must be an accurate, consistent data base; and second, there must be an adequate computer program to retrieve the desired data. Melvin Luthy has been working on the data base, and Robert Stevens has been working on the computer program. At this point, we have 62,500 words in the base written in phonetic transcription with syllabification marked. The transcription system is modified IPA (International Phonetic Alphabet) consisting of the following symbols:

\[
\text{[p, b, t, d, k, q, f, v, θ, ʃ, s, z, ʃ, h, ɣ, m, n, ŋ, l, r, w, j, i, I, e, E, a, u, o, ɔ, ø, aɪ, aʊ, a, ɛ, ɻ, ɜ, ɹ, s, z, ʃ, θ, ɻ, ʒ, ʃ, j, i, ɪ, e, ɛ, a, u, ʊ, ʌ, j, i, r, w, j, i, I, e, E, a, u, o, ɔ, ø, aɪ, aʊ, a, ɛ, ɻ, ɜ, ɹ, s, z, ʃ, θ, ɻ, ʒ, ʃ, j, i, r, w, j, i, I, e, E, a, u, o]}\]

Syllable boundaries are marked with a hyphen. Only segmental sounds are transcribed. Stress is not marked.

Syllabic and postvocalic r is transcribed [ʃ]. Syllable-initial r is transcribed [r]. Hence, the word *encumber* is transcribed [ɪn-kəm-bər], but *encumbering* is transcribed [ɪn-kəm-bərɪŋ].

Syllabic lateral and nasal resonants are transcribed with a schwa [ə] preceding them ([bæt] battle > [bæt-ə]), so that each syllable is justified with a vowel.

Glottal stops are transcribed as the phonemic symbol with which they are associated. For example, the strictly phonetic [bəʔ] *button* is transcribed [bət-ən].

Flapped r's are also transcribed as idealized pronunciation would suggest. That is, the phonetic [lɛr] *ladder* is transcribed [læt-r] *latter*, but [læd-r] *ladder*.

Tense vowels, which tend to be diphthongal, are not transcribed as diphthongs, but rather as [i, e, u, o].

Frequently occurring suffixes such as -ing or ed, are transcribed with preceding consonants as part of the final syllable. This feature enables the user to retrieve more specific ing forms than a trivial listing all words ending in ing, although that, too, is possible, as explained below.

One major problem in entering the data of spoken English is representing the variant pronunciations of a given word. Although we have 62,500 words in the corpus written in English orthography, we have approximately 72,000
correspondences in phonetic transcription. Examples of variant pronunciation entries include the following:

stumbling: stəm-bə-liŋ, stəm-blIŋ
suggestion: sæ-ʃEs-čən, sæ-ʃEs-čən, sæ-g-ʃEs-čən, sæ-g-ʃEs-čən
status: stæt-əs, stet-əs
suspect: sæ-spekt, sæ-spEkt
syndicate: sIn-dr-kət, sIn-də-kət
subject: sæb-ʃEkt, sæb-ʃIkτ
standby: stænd-bai, stæn-bai
strictly: strIkt-li strIk-li
stance: stæns, stænts
caught: kat, kət

The data base has been entered on a Macintosh SE computer on which phonetic characters can be designed. The vocabulary base is a subset (54%) of the current WordPerfect Speller.

Developing a Solution: Computer Program

Robert Stevens' work is focussed on developing a program that will run as fast as possible with the data compressed to fit easily on two floppy discs. It would be nice to display the phonetic notation on graphical display such as the MacIntosh computer provides. The rhymer retrieval system is very versatile, enabling the user to retrieve data in a variety of forms, according to need, as the following examples illustrate:

Perfect Rhyme. Poets speak of "perfect rhyme," which is essentially rhyming with the final vowel of a word, including diphthongs, plus any consonant(s) following it. This type of rhyme is elicited in the program by the function "End of word."

Identical Rhyme. This type of rhyme is the same as perfect rhyme, except the consonant(s) preceding the vowel are also included. This type of rhyme is elicited by the function "last syllable." It retrieves all sounds following the last syllable boundary. In some cases, the last syllable will also be a perfect rhyme.

Word Initial Rhymes. The mirror images of perfect rhyme and identical rhyme are retrievable from the beginnings of
words. That is, the function "first syllable" will retrieve all sounds that occur before the first hyphen. The function "beginning of word" will retrieve the first vowel and any consonant(s) preceding it.

Alliteration. Alliteration is the successive occurrence of a particular consonant as the first sound of successive syllables. Alliterative consonants in specific positions in a word can be retrieved by selecting the desired consonant from the sound menu, and placing it in the desired sound pattern. For example, words with alliterative p's as in peppers or piper can be elicited by entering: p*-p*, where the asterisk represents any sound(s), and the hyphen represents a syllable boundary. The use of an asterisk in this manner is referred to as a "wild card."

Assonance. Assonance is the recurring of a particular vowel sound in successive syllables. Words with desired vowel sounds may be retrieved in the same manner as alliterative consonants were retrieved. For example, if the user were to retrieve words having the vowel [i] in both of the first and second syllables, he could enter: *i*-i*, where the asterisk represents any sound(s) (or lack of sound), and the hyphen is a syllable boundary. This function would retrieve, for example, the word retrieve, and other words with the vowel [i] in both the first and second syllables.

Consonance. Consonance is the repetition at close intervals of syllable-final consonants. The same type of procedure as that used for alliteration and assonance may be used to retrieve repetitive consonants.

Other Sound combinations. It should be clear that with the use of the "wild card" feature, the user may retrieve any consecutive or nonconsecutive sound sequences. This feature makes the rhymer a very useful tool for linguists, teachers, and material developers.

Another appealing feature of the program is that the word on which the cursor is placed may be replaced directly with any of the rhyming words retrieved.

Demonstration of the Program

As part of the demonstration of the rhymer, it seems appropriate that we help "Dear Abby" with her problem of finding other words with the final "gry" sound. A quick check with the rhymer reveals at least the following:

agree, degree, disagree, filigree, pedigree, and vinegary
(There is also another pronunciation of vinegary that includes a schwa vowel between the [g] and the [r])

Some persons have asked for words rhyming with orange.
The rhymer reveals the following "perfect rhymes:

hinge, syringe, astringe, singe, binge, cringe, fringe, tinge, impinge, infringe

(This paper was presented March 14, 1989 in the Deseret Language and Linguistics Symposium at Brigham Young University)
Automatic Recognition of Relative Clauses
with Missing Relative Pronoun

Kenton L. Chauncey
Brigham Young University

As I was pondering my thesis topic a few years ago, Professor Skousen suggested that I do an analysis of the Tagged Brown Corpus. I told him that I wanted to do a thesis in linguistics and that I didn't know anything about performing autopsies. He said it wouldn't really be an autopsy, since the Brown Corpus isn't really dead yet, although it is getting old. So I asked him what the body tag was for. He said it wasn't a body tag; it was grammatical tags. For those of you unfamiliar with the Brown Corpus, it is a corpus of about a million words compiled at Brown University (hence the name) by Nelson Francis and Henry Kučera in the early 1960s. It aims at being representative of American written English and contains 500 samples of about 2000 words each from sixteen different categories, such as newspaper articles and science fiction. The Tagged Brown Corpus has a grammatical tag assigned to each word and each punctuation mark.

There are two big advantages to using a database such as the Tagged Brown Corpus. First, it eliminates the hassle of compiling a corpus. But more importantly, it alleviates the necessity of having to provide a front-end to your program. A front-end performs a preliminary analysis of the data and converts it into some form that your program can use. For example, if you type in a sentence to be parsed, a dictionary lookup will need to be performed to determine the possible parts of speech (or whatever similar convention that your program requires) before parsing can take place. With the Tagged Brown Corpus, the only preprocessing necessary is such as is needed to convert its tags into ones that your program will recognize.

The objective of my thesis was to write a program to locate, in the tagged version of the Brown Corpus, the object relative clauses that do not have an explicit relative pronoun, or, in other words, "zero relatives". Object relatives are always restrictive clauses. Relative clauses with relative pronouns were not considered because the Tagged Brown Corpus tags wh-words according to whether they are used as interrogative pronouns or relative pronouns. So Professor Skousen and I decided that I would search for zero relatives. Below are two object relative clauses; the first with the relative pronoun, the second with the relative pronoun deleted.

(1) The man that I saw was dead.

(2) The man I saw was dead.

The only difference between (1) and (2) is the presence or absence of the relative pronoun. However, there is no difficulty in locating the relative clause in either case. Why?

Now all linguistic theories have some form of handling relative clauses. Most involve rewriting a noun or noun phrase as a noun or noun phrase plus a sentence, such as (3).

---

1 For an expanded version of this paper, see Chauncey (1989).
What I wanted to do was find a simple algorithm to accomplish this in perhaps an ad hoc manner. At least ad hoc in the sense that it was not part of a larger, formal linguistic theory. The rationale was that we process language in some sort of left-to-right fashion (left-to-right from the way we read English). That is, when we listen to someone speaking, we don’t wait until they have finished speaking to begin processing. We process on the fly, taking what is being said and incorporating it into what has already been said, and making predictions on what will be said. Each word that we hear adds another piece to the puzzle, confirming or disproving our predictions of what is to come. In the case of relative clauses, we normally have a relative pronoun to indicate the beginning of the clause. Although relative pronouns can also be used as interrogatives, they usually occur in mutually exclusive environments. But what about zero relatives? We can recognize the relative clause just as easily in “the man I saw” as we can in “the man that I saw”. In these relative clauses, the relative pronoun must be redundant. So the question was, could we take advantage of this redundancy and find a simple way to locate zero relatives?

To digress briefly, you will probably agree that language is the most complex activity that humans engage in on a daily basis. Fortunately, language is also extremely redundant. That is, not every sound, letter, word, or phrase is necessary for the hearer to understand the message.

Some parts of the message can be predicted by other parts that are present. It is this redundancy, ironically, that makes language a useful medium for communication. With zero redundancy, every bit of information would be completely necessary. The hearer would need to hear every sound uttered by the speaker. To miss even one sound, whether by inattention, noise, etc., would disrupt the integrity of the message. It has been estimated that English, and perhaps all languages, is 50% redundant. At any rate, there is much redundancy, and this should be considered and taken advantage of when trying to parse natural language.

So equipped with this hope of redundancy, I set out to break the code of zero relatives. As I thought about my thesis topic, it seemed alternately trivial or impossible. I had collected a variety of relative clauses, and all the zero relatives seemed to involve adjacent noun phrases. It seemed that all I needed to do was to locate all of the noun phrases, determine which were adjacent, and collect my $200 as I passed GO.

Then I started studying the various possibilities of adjacent noun phrases, such as indirect object/direct object (“He gave me the ball”), verbs that take two objects (“They elected Bush President”), verb complements (“He told me it wasn’t ready”), subjects preceded by introductory prepositional phrases (“Of course it’ll be dark when I get home”), etc., etc.—not to mention the adverbial constructions that can go almost anywhere. It seemed that the only analysis I would get would be psychiatric.

However, it is also possible to overestimate the difficulty of a problem. Whenever there is one simple way to solve a problem, there are usually numerous more difficult ways to solve the same problem. Albert Einstein, who had a knack for finding the lowest common denominator, said: “Everything should be made as simple as possible, but not simpler.” (Minsky, 1986, p. 17) The simplest method was in the data of the Brown Corpus all along. It was a surprising discovery to me that it could be so simple.
But as the creator of Transformational Generative Grammar asserts, "It is important to learn to be surprised by simple things" (Chomsky, 1988, p. 43).

When I began, my goal was to locate 100 percent of the zero relatives and to flag none that were not zero relatives, which corresponds to Chomsky's definition that a grammar "generates all of the grammatical sequences...and none of the ungrammatical ones" (Chomsky, 1969, p. 13). However, I soon realized that success is often measured (and, more particularly, would have to be in this study) at something less than perfection. I began to envision the program more as a component of a larger analysis routine, a heuristic subroutine that would flag occurrences with reasonable certainty, which could then be verified or rejected by a more exacting syntactic analysis in the larger routine.

But even shooting for less than 100 percent seemed difficult. For example, my initial algorithm required that the program locate noun phrases and determine if they were "close" to each other. But besides determining what, if anything, could separate the head noun from the subject of the relative clause, how could the program differentiate between zero relatives and other examples of noun phrase + noun phrase? How would it deal, for example, with indirect object/direct object, two objects, and verb complements? At this point, I was ready to call in the cavalry.

About this time, I began working with actual tagged data. Almost immediately it was apparent that step 1 above, locating noun phrases, was impossible--at least in the sense that it was far beyond the scope of a Master's thesis. The first sentence in the Brown Corpus will serve as an example. (4) below shows the sentence with the respective tags placed below each word. Table 1 contains some of the more frequently used tags.

(4) The Fulton County Grand Jury said Friday an
AT NP-TL NN-TL JJ-TL NN-TL VBD NR AT
investigation of Atlanta's recent primary election
NN IN NPS JJ NN NN
produced "no evidence" that any irregularities took place.
VBD AT NN CS DTI NNS VB NN .

Ignore the hideous first noun phrase—which is AT NP-TL NN-TL JJ-TL NN-TL. Temporarily assume that all tags ending with "-TL" (signifying "title") will collapse into a single proper noun (although this assumption does not hold up). The second noun phrase is simply the single word "Friday", which is tagged as an adverbial noun. It falls between a verb and an article, so there is little trouble recognizing it as a separate unit. The third noun phrase, "an investigation", is also relatively easy to separate, since it follows the standard AT + NN (article + noun) formulation, and is delimited on the right by a preposition.

---

2See Francis and Kučera (1979) for a complete list of tags.
Table 1
Some Frequently-Used Tags

<table>
<thead>
<tr>
<th>TAG</th>
<th>DEFINITION</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>article</td>
<td>a, an, the</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td>two, 2</td>
</tr>
<tr>
<td>IN</td>
<td>preposition</td>
<td>to, in</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>large</td>
</tr>
<tr>
<td>NN</td>
<td>singular noun</td>
<td>man</td>
</tr>
<tr>
<td>NNS</td>
<td>plural noun</td>
<td>men</td>
</tr>
<tr>
<td>NP</td>
<td>proper noun</td>
<td>Bill</td>
</tr>
<tr>
<td>PPO</td>
<td>obj. personal pronoun</td>
<td>me, him, them</td>
</tr>
<tr>
<td>PPS</td>
<td>3rd sing. nom.</td>
<td>he, she, it, one</td>
</tr>
<tr>
<td>PPSS</td>
<td>other nom. personal pronouns</td>
<td>I, we, you, they</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base form</td>
<td>sing</td>
</tr>
</tbody>
</table>

Now the plot thickens. Following the preposition is a possessive proper noun. It is possible that this noun could be the object of the preposition and therefore constitute a noun phrase. However, an adjective follows, pointing to another noun farther down the line. Does the possessive go with this coming noun or does it stand alone? To further complicate matters, there is not just one noun following the adjective but two nouns, back-to-back, and a finite verb following the second noun. So do we break after the possessive? Or do we continue on and break between the two nouns, assuming, perhaps, that the second noun is a mass noun (therefore not requiring an article) and the subject of the verb “produced”? (Do we now have to check for mass nouns, which will then need to be included in a dictionary?) Or do we somehow correctly realize that the entire construction, i.e., “Atlanta’s recent primary election”, forms a single and moderately complex noun phrase?

The situation does not improve; rather, it gets rapidly and exponentially worse. Practically every sentence is a parsing nightmare. Ain’t natural language grand!

Since it was too late to change topics, I continued to analyze my Brown subcorpus by hand (or eye), hoping for some kind of breakthrough. Soon I had a few zero relatives to assess. The first one that I found is contained in (5) below.

(5) The largest hurdle the Republicans would have to face is

AT JJT NN AT NPS MD HV TO VB BEZ

a state law which says that before making a first race,

AT NN NN WDT VBZ CS CS VBG AT OD NN ,

one of two alternative courses must be taken:

CD IN CD JJ NNS MD BE VBN :

The deleted "that" occurs between the two noun phrases "the largest hurdle" and "the Republicans", the tags for the noun phrases being "AT JJT NN" and "AT NPS", respectively. The next zero relative was contained in the segment "the highway bond money it approved". And others followed. As I looked at them, a light finally flickered.
In every case, a noun (either a "NN" or some variant thereof) was immediately followed by an "AT" or a personal subject pronoun. It couldn't be that simple, could it? Well, for my purposes, did I really care what the entire noun phrase looked like? No, I was simply trying to locate the deleted "that". So the question of whether it could be that simple to locate the deleted "that" was an empirical question.

I continued analyzing my test corpus. Seven of the first 10 zero relatives were either NN + AT (that is, singular noun + article)\(^3\) or NN + PPS(S) (singular noun + nominative personal pronoun). Seventy percent with just three checks! English has a much greater diversity with relative clauses than that, but it seemed a promising start. However, finding the correct zero relatives was only half of the battle. Overflagging had to be considered next.

Checking overflagging would be a task ideally suited for the computer. I hurriedly completed the test program. It allowed me to search for various adjacencies, such as NN + AT. Holding my breath, I ran the program on the first part of the corpus that I had already gone over. It worked! The program found the three NN + AT that were zero relatives and only found three occurrences of NN + AT that weren't zero relatives. It wasn't 100 percent accurate, but the percentage was well within tolerance. NN + PPS and NN + PPSS worked even better. All four zero relatives were located and there wasn't a single overflagging. It seemed now that much of the problem would be discovering the potential adjacencies, determining the degree of their productivity, and deciding which adjacencies to incorporate into the program.

Once I accepted adjacency as the primary test for zero relatives, I also accepted as the primary assumption (and hope) that there would not be any checks that would occur frequently enough that they would have to be included in the program, but at the same time cause so much overflagging that they would skew my results. This assumption did hold up, but only tenuously in some cases.

The above are just a few of the many adjacency pairs that were examined. Some adjacencies were accepted or rejected on the basis of testing 100 sentences, either by overwhelming evidence or by "Sprachgefühl". Some adjacencies were tested through half of the Brown Corpus. For example, "Mr." and "Mrs." and "Miss" are tagged as NP. As shown in Table 2, NN + NP is not a productive pair. However, if the NP is "Mr.", then the pair is productive (5-6), although very infrequent. "Miss", on the other hand, is neither frequent nor productive (1-3). NN + NP(Miss), therefore, was not included.

After extended analysis, I began the crucial testing phase. The main test that the program was to be subjected to was the analysis of a 10,000 word subcorpus (the "Test Corpus") to be chosen by Prof. Skousen. I would first go through the corpus and locate the zero relatives manually. The program would then have to find 90% of the zero relatives. It was allowed an overflagging margin of 200%.

\(^3\)The "+" used between tags is not meant to be construed as the adjoinment symbol used in phrase structure grammars. Rather, it is more a symbol of concatenation or adjacency.
I analyzed the first Test Corpus and ran the program on it. To my delight, the program worked on the very first try. However, a minor error had nearly escaped detection. Each mark of punctuation counts as a separate record in the tagged Brown Corpus. When I searched through to find out exactly how many punctuation records there were, there were more than I had estimated, making the Test Corpus somewhat less than 10,000 words. This did not seem to pose a problem. I would simply add the additional text, analyze it, run the program on it, and begin my writeup. Had I restricted the text to 10,000, this is what would have happened. However, I wasn't worried about a few extra words and ended up testing about 10,500 words. In those last 500 words, against the odds, there were two zero relatives that my program could not find. This dropped my percentage to just below 90%.

Rather than chop off the last 500 words, I went through another Test Corpus analysis. This, and the final third Test Corpus analysis, was actually unnecessary. In my haste to begin the second analysis, I failed to notice that one of the zero relatives that the program had not found was not actually a zero relative. (It was a verb complement construction.) This pushed the percentage found to just over 90%. However, I didn't notice this error until after I had already performed the other two tests. Now that it is done, I am glad that I was forced to continue. Two out of the three Test Corpora were successful. And 30,000 words provide a much better test of the algorithm than one 10,000 word test.

There were 78 checks in the final version of my program. Of these, 18 different individual tests were triggered to locate zero relatives, as shown in Table 3. Of these 18, six did not find any zero relatives. So with 12 checks, the program located 89% and overflagged only 79%.  

### Table 2

**Some Potential Adjacencies and Their Productivity**

<table>
<thead>
<tr>
<th>TEST</th>
<th>number of sentences (approx.)</th>
<th>found</th>
<th>overflagged</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN + AT</td>
<td>2000</td>
<td>7</td>
<td>40</td>
<td>&quot;loan the&quot;</td>
</tr>
<tr>
<td>NN + PPS(S)</td>
<td>2000</td>
<td>16</td>
<td>10</td>
<td>&quot;society we&quot;</td>
</tr>
<tr>
<td>NNS + AT</td>
<td>2100</td>
<td>3</td>
<td>28</td>
<td>&quot;changes the&quot;</td>
</tr>
<tr>
<td>CD + AT</td>
<td>5000</td>
<td>1</td>
<td>12</td>
<td>&quot;one the&quot;</td>
</tr>
<tr>
<td>NN + NP</td>
<td>1500</td>
<td>2</td>
<td>30</td>
<td>&quot;death March&quot;</td>
</tr>
<tr>
<td>NN + NN-TL</td>
<td>3500</td>
<td>2</td>
<td>16</td>
<td>&quot;way Mother&quot;</td>
</tr>
<tr>
<td>NN + NN</td>
<td>1000</td>
<td>0</td>
<td>48</td>
<td>&quot;primary election&quot;</td>
</tr>
<tr>
<td>NN + JJ</td>
<td>1000</td>
<td>0</td>
<td>17</td>
<td>&quot;state deaf&quot;</td>
</tr>
<tr>
<td>PPO + AT</td>
<td>3000</td>
<td>0</td>
<td>17</td>
<td>&quot;you a&quot;</td>
</tr>
<tr>
<td>PPO + PPS(S)</td>
<td>10,000</td>
<td>0</td>
<td>2</td>
<td>&quot;it the&quot;</td>
</tr>
<tr>
<td>PPS + AT</td>
<td>26,000</td>
<td>5</td>
<td>6</td>
<td>&quot;statement Mr.&quot;</td>
</tr>
<tr>
<td>NN + NP(Mr.)</td>
<td>26,000</td>
<td>1</td>
<td>3</td>
<td>&quot;diva Miss&quot;</td>
</tr>
<tr>
<td>(Miss)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3
Analysis of the Combined Test Subcorpora

<table>
<thead>
<tr>
<th>TEST</th>
<th>found</th>
<th>over flagged</th>
<th>missed</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN + PPS</td>
<td>23</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>NN + AT</td>
<td>5</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>NN + PPSS</td>
<td>26</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>NNS + PPS</td>
<td>12</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>NN + PP$^*$</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>NN + PPSS$^*$</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NNS + PPSS$^*$</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>NNS + PPSS</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>PN + PPS</td>
<td>4</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PN + PPSS$^*$</td>
<td>2</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>NNS + PN$^*$</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NNS + PPSS$^*$</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>CD + PPS</td>
<td>2</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CD + PPSS</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CD + AT</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NN-TL + PPS</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NN-TL + AT</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NN(time) + NP</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NNS + AT</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>NP + NP</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>NN + ABN</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>NN + DTI</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>NN + NN</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>NN + NP</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>NNS + PPSS$^+$</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>PN + EX</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>PPO + PPSS$^+$</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>TOTAL</td>
<td>89</td>
<td>79</td>
<td>11</td>
</tr>
</tbody>
</table>

In retrospect, there were at least two good reasons for not trying to find 100% of the zero relatives. One was purely practical, one theoretical. The practical reason involves the tradeoff between constructions found and constructions overflagged. Since every adjacency test involved some overflagging, increasing the number of tests in order to find 100% of the zero relatives would also increase the number of overflaggings. At some point, the law of diminishing returns becomes too great and the number of overflaggings too high. The 90% goal that we set seems to have been about right.

From a theoretical standpoint, the model that we use may determine whether we believe that 100% is possible. According to the structural vs. analogical dichotomy presented in Skousen (1986), the structural approach relies on rules to predict behavior and assumes that a minimal number of rules, perhaps a single rule (the rule), can cover the behavior in question. An analogical approach, however, believes in a multiplicity of factors (including redundant factors) that work together to predict behavior, one set of factors being used in one situation, another set being more useful in another situation,
and so on. The present study seems to support the latter approach if it turns out that we, as humans, use similar strategies to process language as those contained in my program.

The final task for my program was to search the entire Brown Corpus for occurrences of multiply-embedded zero relative clauses. For this search, I added most of the checks that had found any zero relatives at all. The ones that I didn't add were those with only one occurrence in half of the Brown Corpus versus numerous overflaggings. I then arranged that every sentence that contained two or more zero relatives be printed out. I chose not to try to constrain the distance between the relatives in any way, for fear of having a multiple-embed fail through the cracks. There was one example of a two-deep relative clause in over 25 pages of sentences with two zero relatives. The sentence is (6) below.

(6) The card the man I was shadowing had filled out was
AT NN AT NN PPSS BEDZ VBG HVD VBN RP BEDZ
still on the counter.
RB IN AT NN .

So far as I have been able to tell, this is the first sentence with multiple nested embedding of this type ever mentioned in a linguistic treatise that was an actually-occurring sentence. It was written by A. A. Fair, which was a nom-de-plume of Erie Stanley Gardner.

I was surprised to locate an occurrence of this construction. In my prospectus I predicted that no examples would be found, because I didn't believe that any examples existed (my "two deep is too deep" hypothesis). Although I still contend that this construction is too difficult to process in order to be counted as a "grammatical" sentence, the fact is that it exists, and it cannot just be swept under the carpet. One of the great things about working with real data is the possibility of confirming or disconfirming a hypothesis.

And so, with apologies to Erie Stanley Gardner, I will profess to have solved "the case of the missing relative".

---

4See Appendix H in Chauncey (1989) for a list of multiply-embedded zero relatives.
Bibliography


Computer Analysis of Aymara Morphology:  
A Two-Level, Finite-State Approach

Kenneth R. Beesley  
Stuart N. Newton  
ALPNET  
190 West 800 North  
Provo, UT 84604  
(801) 379-2300

ABSTRACT

This paper describes a working computer program which performs morphological analysis and dictionary lookup of words in Aymara, an Andean language spoken in Bolivia and Peru. Aymara is noteworthy for the richness of its morphology; there are, for example, 24 suffix classes that can follow a verb root. The analysis is written using a morphology analysis tool based on Koskenniemi's Two-Level Theory of morphology; this generalized morphology tool is driven by lexicons and tables and can be applied to other languages, including Semitic languages, without modification of the code.

INTRODUCTION

The Aymara language, spoken by three million people in Bolivia and Peru, is well known for its highly complex and productive morphology. Noun roots, for example, can theoretically exhibit up to 13 suffixes chosen from 13 different suffix classes; verb roots can be followed by suffixes from 24 different suffix classes. About 200 individual suffixes are available. Possibilities for nominalization of verbs and verbalization of nouns further complicate the picture.

In a recent project at ALPNET, we undertook to write a program to automatically analyze the morphology of Aymara words written in a standardized orthography. Our goal was to identify each possible reading of a written word and to show, for each reading, the breakdown of the morphemes, a rough translation for each morpheme, and an overall characterization of the word in terms of features like tense, number and gender.

The traditional automated approaches to morphological analysis of written language involve "cut and paste"
operations whereby suspected affixes are cut off until a suspected stem is left. Then that stem is looked up in a dictionary. For a morphologically complex language, the cutting and looking-up steps must often be repeated several times, perhaps in different orders, before a successful "reduction" occurs. In addition, phonological or graphological changes may occur and be reflected in the surface orthography; this may require various kinds of compensatory "pasting" on the suspected stem, and even on the affixes, before they can be successfully identified or looked up. It is almost inevitable for such cut-and-paste approaches to become large, difficult to maintain and improve, and completely ad hoc.

Two-Level Morphology, developed by Kimmo Koskenniemi (1983, 1984), Lauri Karttunen (1983) and others (Karttunen & Wittenburg, 1983; Gajek et al., 1983; Bear, 1988), is a powerful, elegant, and increasingly popular theory for generating and analyzing natural-language morphology. With some clever techniques for using lexicons (i.e. dictionaries) to drive the analysis and generation processes, computer implementations of Two-Level Morphology are not only theoretically interesting but highly robust, thorough, and efficient. In October and November of 1988, we successfully implemented Koskenniemi’s two-level theory as a generalized morphological analysis tool. Then in January and February of 1989, we built lexicons and other tables to allow the tool to correctly analyze a large sample of Aymara words. We continue to expand and improve the system from time to time.

TWO-LEVEL MORPHOLOGY

Two-Level Morphology is so named because it postulates two distinct but interrelated levels of representation for words, in our case a slightly abstract and regularized lexical level and a surface orthographical level.

Lexical-Level
Surface-Level

The lexical level for any valid Aymara word is a concatenation of morphemes, in a legal sequence, each morpheme consisting of a string of characters as they appear in the lexicon. The surface level is a string of characters representing the word as it actually appears in our standardized surface orthography. By convention, the lexical level is always written above the surface level. The following noun example utaninaca shows a lexical string and its surface "realization"; the plus and pound signs at the lexical level represent word and morpheme boundaries which are never realized as such on the surface.
Lexical: #uta+ni+naca#
Surface:  uta ni naca

# = word-boundary
uta = house
+ni = owner of
+naca = plural
# = word-boundary

(trans. "owners of the house")

The following verb example lurana means "to make/do."

Lexical: #lura+Na#
Surface:  lura Na

# = word-boundary
lura = make/do
+Na = infinitive
# = word-boundary

It should be emphasized that in two-level theory, the lexical string is not transformed or changed into the surface string; both levels exist simultaneously and are related with morphological rules. Relation of the two levels is not always so trivial, as in the following example, also based on the root lura, where the final a in lura is not realized at the surface level.

Lexical: #lura-i+wa#
Surface:  lur i wa

# = word-boundary
lura = make/do
-i = 3rd person present
+wa = primary emphasis
# = word-boundary

(trans. "he/she makes/does")

In Aymara, as in most languages, a single string of surface characters may be a valid realization of several different lexical strings. In some languages, especially Arabic and Hebrew, which have optional realization of short vowels, a single lexical word may have many possible surface realizations.

The alphabet used in two-level theory is a set of character pairs, called "concrete pairs." Each concrete pair consists of a lexical character and a surface character which is one of its possible surface realizations. Linguists must discover the concrete pairs required for each language and declare them for use by the analysis program. Concrete pairs are represented with the lexical character above the
surface character, or, especially for computer work, with the sequence Lexical_Char:Surface_Char. A reserved null or empty character, here zero, can be used to show that a lexical character is not realized at all on the surface ("zero realization"). The following concrete pairs figure in the luriwa example shown above.

Vertical Representation:

Lexical:       #lura-i+wa#
Surface:      0lur0010wa0

Horizontal Representation:

#:0           "lexical pound sign realized as zero (nothing)"
1:1           "lexical 1 realized on the surface as 1"
u:u           "lexical u realized on the surface as u"
r:r           "lexical r realized on the surface as r"
a:0           "lexical a realized as zero (nothing)"
a:a           "lexical a realized on the surface as a"

etc.

Where multiple options for realization exist, as for lexical a, which can be realized as surface a or as nothing ("zero"), morphological rules control the realizations by referring to lexical and surface context. In other words, morphological rules resolve and control the discrepancies between the lexical and surface levels. For example, the following is a rule to allow and force the realization of lexical a as surface 0 if and only if it is in the context preceding a lexical minus sign (which is always realized as zero).

a:0  <=  __  -=:0

This rule comprises two logical implications. First, the left-arrow or "if" part of the rule specifies that if lexical a appears in the indicated two-level context, then it must be realized on the surface as zero. In other words, lexical a is forced to be realized as nothing in this context, and the rule will block any other realization of lexical a in this context.

Lexical: a-
Surface: 00         Required realization of a:0
                  in this context

Lexical: a-
Surface: a0         Illegal realization of a:a
                  in this context
The second or "only if" part of the rule, indicated with the right arrow, specifies that lexical a can be realized as nothing ONLY IF it occurs in the indicated context. That is, any attempt to realize lexical a as zero when it is not followed by a lexical minus sign will be blocked by the rule.

Lexical: a-
Surface: 00 Allowed realization of a:0 in the specified context

Lexical: a+
Surface: 00 Illegal realization of a:0 in a different context

Zero realization of short a before a strong suffix is generalizable to i and u as well, resulting in the following rule.

a:0, i:0, u:0 <=> _ -:0

Depending on the morphological complexity and degree of regularity in a language, dozens of such rules may be necessary to correctly relate the lexical and surface levels. About twenty rules were required for Finnish in Kimmo Koskenniemi’s original project. We anticipate that about forty rules will ultimately be required for Arabic, another language we are working on. Aymara, which is morphologically rich but surprisingly regular, appears to require only four rules; while Esperanto, which is completely regular, required no morphological rules at all.

CONTROL MECHANISMS

In the ALPNET implementation of Two-Level theory, as in others, lexicons drive the search process while the rules serve merely as filters. Developers create a separate sublexicon for each morpheme class (such as the 24 classes of verb suffix), and these lexicons are stored and accessed as letter-transition trees. For example, roots might be listed in an "unloaded" lexicon as shown below. Using traditional ALPNET conventions, a percent sign is used to separate entries, and n and v indicate the noun and verb categories respectively. Linguists could create such a lexicon specification using any word processor.
There is also a class of morphemes in Aymara, here called $v_I$, which can appear directly after a verb.

A lexicon-loader program automatically organizes these entries into letter-transition trees like the following, where the period represents the root of a tree, and parentheses enclose the information belonging to a lexical item, including the category and the translation (only the category is shown here).
For analysis, the program is given the information that a word in Aymara must begin with a root morpheme; and it will explore any lexical paths through the root lexicon that the surface word, the concrete pairs, and the morphological rules will support. Analyzing the surface word wararaNa, for example, the program will successfully trace a lexical path through the root lexicon to find wara with a reading of category "v" for verb. That is, it will start searching at the root of the letter-transition tree, trying to match the surface letters of the input string with the lexical letters in the tree. Because of trivial lexical-to-surface correspondences such as w:w, a:a, and r:r, the program will eventually discover the potential lexical morpheme wara, leaving the surface letters raNa still to be accounted for.

Once the verb morpheme wara is found, the system then consults a table of "continuation classes" to see what kind of morpheme can follow a verb. The table entry will look like the following, where v1 through v22 are names of other lexicons, each containing a class of verb suffixes.

("v" v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 v13 v14 v15 v16 v17 v18 v19 v20 v21 v22)

This table entry tells the system that a verb root can (and, indeed, must) be followed directly by a morpheme
chosen from any one of the indicated suffix classes. The system then searches each of the indicated lexicons in turn, trying to account for the rest of the letters in the surface word. In this case, the system first looks in the v1 lexicon and succeeds in finding the +ra morpheme, which means "undo" and has the category "vI." Because the concrete pairs will include +:0 ("the plus sign is realized as nothing on the surface"), the program will be able to follow a path through lexical +, lexical r and lexical a even though the surface word contains only r and a.

The continuations for a "vI" morpheme like +ra are the following:

("vI" v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 v13 v14 v15 v16 v17 v18 v19 v20 v21 v22)

The +Na suffix, indicating the infinitive, will eventually be found in the v22 lexicon, completing the analysis. The continuation class for a v22 will indicate that it can validly end a word.

("v22" v23 neut end)

When it succeeds, the system prints out the path of lexical letters that it found while traversing the lexicon trees. It also displays the collected lexicon translations for each morpheme.

Lexical: #wara+ra+Na#
Surface: wara ra Na

# = word-boundary
wara = pour
+ra = undo
+Na = infinitive
# = word-boundary

This reading for wararaNa, which might be rendered as "to unpour," is semantically questionable at best--pouring is the kind of activity that normally cannot be undone. The program, exploring all paths, also finds another +ra morpheme, meaning "distributive," in the v2 dictionary. This path yields a second, preferable reading.

Lexical: #wara+ra+Na#
Surface: wara ra Na

# = word-boundary
wara = pour
+ra = distributive
+Na = infinitive
# = word-boundary
Each potential morpheme that is identified will have a category and an appropriate set of continuation lexicons to explore next. Typically in Aymara, an entire surface word will require successful searching through several lexicons before the entire surface word is accounted for and the lexicons "bottom out" with a lexical solution. Sometimes, as for wararaNa, multiple paths through the lexicons will bottom out, resulting in multiple analyses. The goal, of course, is to generate all of the valid analyses and no invalid ones.

Many possible paths that are allowed by the continuation classes soon run into a lexical dead end. The program then goes back to explore other possible tree paths. Because the trees are searched with recursive routines, all indicated paths eventually get explored, making the system very thorough. It should be noted that because the lexicons actually drive the analysis, any analysis paths that are not supported by the lexicons are immediately abandoned.

The strength and power of a two-level analysis is in its ability to take a string of surface characters as input and follow a path of lexical characters through the lexicon, even if there are, because of phonological effects or graphological conventions, considerable discrepancies between the two strings. Given the surface string luriwa, for example, the program will tentatively follow a path in the root lexicon through lexical l, lexical u, lexical r, and lexical a, finding the root lura. In other words, the system automatically finds lexical lura even though it has only lur to work with on the surface. This is possible because the linguist tells the system that a lexical a can be realized as nothing on the surface. The morphological rules, including those controlling the realization of lexical a, run in parallel with the recursive tree searching; they will ensure, in this case, that the present context allows lexical a to be realized as nothing.

At no time does the system resort to ad-hoc "cut-and-paste" methods to reduce surface words down to a root that can then be looked up. In fact, there is never any separate lookup step at all; readings for the various morphemes are simply collected along the path to a successful analysis.

LINGUISTIC METHODOLOGY

The first step in analyzing the morphology of Aymara, or any other language, is to declare the surface alphabet that will be used by the program. There are several roughly phonemic surface orthographies used to write Aymara, some
of them not completely satisfactory because they are influenced rather arbitrarily by standard Spanish orthography. The velar stop or /k/ phoneme, for example, is often represented as c before a and u but as qu before i. The /i/ phoneme is represented as e in some completely predictable contexts, indicating the pronunciation of the allophone from a Spanish point of view.

Reflecting the limitations of our character set on the Macintosh, the desire to handle all the commonly used surface orthographies, and the recognition that a more consistent phonemic orthography would simplify rule writing, we have defined our own standard surface orthography for Aymara. Other orthographies can be mapped mechanically and quite trivially into our standard by means of simple character replacement tables. We have used N for the ñ; J for the velar fricative commonly spelled with the digraph jji; a, i, and u for the short vowels; and A, I, and U for the long vowels. The /k/ phoneme is consistently written with c, and k represents the uvular stop. We continue the common practice of using digraphs for ejective consonants such as in p’.

The second step is to identify and classify all of the productive affixes in the entire language. Aymara morphology uses only suffixes, but the wealth of suffixes can be staggering, as in the following example.

**Lexical:** #uta+ma+nca+pJa+samacha-i+wa#
**Surface:** uta ma nca pJa samach i wa

- # = word-boundary
- uta = house
- +ma = 2nd person possessive (your)
- +nca = in (verbalizer)
- +pJa = plural
- +samacha = it appears
- -i = 3rd person present tense
- +wa = primary emphasis
- # = word-boundary

(trans: "it appears that they are in your house")

Fortunately for us, the *Rudimentos del idioma Aymará*, published originally by the Canadian Baptist Mission (1956), provided very helpful lists showing 23 classes of verb suffix, 12 classes of noun suffix, and a neutral class of morphemes that can attach to nouns or verbs. We started by taking each morpheme class and creating a separate sublexicon.

The third task is to specify the order in which morphemes can occur. Again, we were fortunate that the *Rudimentos* had already specified the relative ordering of all the morpheme
classes. We adopted the obvious convention of naming the verb suffix classes "v1" through "v23" and the noun suffix classes "nl" through "n12." In the simplest view, an Aymara word consists of a root and a selection of suffixes optionally chosen, in the specified order, from the appropriate suffix classes. It is roughly analogous to a Chinese restaurant where you can choose one item from Column A, one (or none) from Column B, one (or none) from Column C, etc.

Inside the program, the morpheme sequencing possibilities are indicated in the continuation classes. For each lexicon category, the continuation table indicates the names of the sublexicons in which the search can continue. Each possible continuation path will be fully explored until the program succeeds, runs into a lexical dead end, or is blocked by a violated rule.

The work was not completely done for us, however. There were of course numerous errors and omissions in the charts. In addition, the linguist must specify which morphemes can validly end a word. Proper specification of the continuation classes remains a technical and creative linguistic task which requires study, experimentation with the program, and perhaps work with a native informant.

The final task is to account, via morphological rules, for the various discrepancies between the lexical and surface levels. One linguistic challenge was to handle "strong" and "weak" suffixes in Aymara. In traditional terms, strong suffixes cause the phonetic and orthographic deletion of a short vowel immediately preceding the suffix. The combination of lura, meaning "make" or "do," with the weak suffix Na, for example, causes no vowel deletion.

\[
\text{lura Na} \\
\text{luraNa} \quad \text{(trans. "to make/do")}
\]

The combination of lura with the strong i suffix, however, causes deletion of the preceding a in lura.

\[
\text{lura i wa} \\
\text{luriwa} \quad \text{(trans. "he/she makes/does")}
\]

Viewing the language diachronically, it is quite probable that at one time some overt phonetic criterion naturally correlated with strongness or weakness. Synchronically, however, no such criterion is discernable, as is clearly evidenced by the existence of a strong suffix pronounced /ta/ and a weak suffix also pronounced /ta/ (from a separate sublexicon). Aymara learners, and our system, simply have to be told whether a suffix is strong or weak.
We adopted the arbitrary convention of preceding each strong suffix at the lexical level with a minus sign, which seemed an appropriate mnemonic for deletion. While it is not formally necessary, we precede weak suffixes, at the lexical level, with a plus sign; this merely facilitates the reading of lexical strings.

The two-level representation of the lura examples, repeated here, shows the different effect of strong and weak suffixes on the preceding vowel.

<table>
<thead>
<tr>
<th>Lexical:</th>
<th>Surface:</th>
</tr>
</thead>
<tbody>
<tr>
<td>#lura+Na#</td>
<td>lura Na</td>
</tr>
<tr>
<td>a:a before a weak suffix</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lexical:</th>
<th>Surface:</th>
</tr>
</thead>
<tbody>
<tr>
<td>#lura-i+wa#</td>
<td>lur i wa</td>
</tr>
<tr>
<td>a:0 before a strong suffix</td>
<td></td>
</tr>
</tbody>
</table>

In two-level terms, the presence of a strong suffix does not cause deletion but rather zero realization of a preceding lexical vowel.

Note that the lexical morphemes include the plus or minus sign as part of the lexical entry. That is, the +Na suffix consists of three lexical letters, and the -i suffix has two lexical letters. Morpheme boundary characters, such as the pound sign, the plus sign and the minus sign, never appear as such on the surface, but they can affect the realization of other letters around them. The zero-realization rule, repeated here, mathematically "knows" when a strong suffix follows a lexical vowel because it looks for the characteristic minus sign.

\[ a:0, i:0, u:0 \Rightarrow -:0 \]

Another Aymara phenomenon to account for is verbalization, whereby a noun root (or a noun root with some noun suffixes on it) can continue on with verb suffixes. The overt manifestation or surface realization of such verbalization is the lengthening of the final vowel in the noun. For example, the noun root jaki means "person"; when the final vowel i is lengthened, indicated in our orthography as I, the word is verbalized and can continue with a verbal suffix like -ta. We propose a verbalizing morpheme arbitrarily spelled & at the lexical level.
Lexical: #jaki&-ta+wa#
Surface: jakI ta wa

# = word-boundary
jaki = person
& = verbalizer
-ta = 2nd person present
+w = primary emphasis
# = word-boundary

(trans. "you are a person" or, more literally, "you person")

The two-level rule to account for this phenomenon is the following:

a:A, i:I, u:U <=> _ &:0

As with the minus sign, the ampersand at the lexical level is never realized as such on the surface, but it affects, through the rules, the realization of lexical vowels preceding it. A small number of Aymara suffixes, which we have termed "super suffixes," begin with a lexical ampersand.

MORPHOLOGY RULES

It is not appropriate or possible here to go into a detailed explanation of how morphology rules work in a two-level system; this information is available elsewhere (Koskenniemi, 1983; Karttunen, 1983). It should be pointed out, however, that the rules do not drive the analysis but simply provide an additional filtering at each step of the path through the lexicon trees.

Rules are compiled into highly efficient finite-state transducers that mathematically keep track of context and realization restrictions by reacting--moving into different states--every time a lexical transition, i.e. a new concrete pair, is proposed. All the rules operate in parallel. Whenever a transition causes any one of the rules to move into an illegal state, this signifies that the rule has been violated, and that search path through the lexicons is immediately blocked.

The following diagram shows the compilation of the verbalization rule. By convention, the machine starts in state 1, and 0 is the failure state. The symbol -A stands for the complement of A, that is, any letter in the alphabet other than A. The equal sign stands for any letter, or, by convention, any letter not designated more specifically elsewhere in the rule. Exit states, those states which can legally end a word, are indicated with double circles.
These spaghetti diagrams are converted to equivalent transition matrices for use by the program.

\[
\begin{align*}
A & = \begin{bmatrix}
1 & 2 & 2 & 2 & 1 & 3 & 3 & 3 & 1 \\
2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
3 & 2 & 2 & 2 & 0 & 3 & 3 & 3 & 1 \\
\end{bmatrix}
\end{align*}
\]

The other two rules required so far for Aymara are the following:

Realization of lexical h as zero when it is preceded by k:k and followed by a vowel realized as zero and a strong suffix

\[h:0 \iff k:k \quad v:0 \iff 0\]
Realization of lexical J as zero when it is preceded by p:p and followed by a vowel realized as zero and a strong suffix that begins with c

\[ J:0 \iff p:p \_ v:0 \_ :-0 \ c:c \ _h:_h \]

The ability to account for the lexical-surface discrepancies with only four rules reflects the traditional observation that Aymara is almost supernaturally regular.

CRITICISMS AND ENHANCEMENTS

For practical purposes in computer analysis of natural-language morphology, we have been extremely pleased with our two-level linguistic tool. However, the theory or perhaps some implementations of it have practical limitations and are subject to various theoretical criticisms. First, the two-level theory can be criticized for allowing the linguist to inject whatever characters he or she wants or needs into the rather abstract lexical level. The danger, of course, is that a linguist will simply kludge the lexical level in whatever way is necessary to account for the surface strings. For Aymara, we have helped ourselves to lexical word-boundary characters and morpheme-boundary characters. The morpheme-boundary characters, +, −, and & are used as a kind of feature to mark suffix types: weak, strong, and super. We have employed similar markers in our analysis of Arabic. For practical work, we fail to see such abstractness at the lexical level as a serious weakness.

Barton (1987) has shown mathematically that two-level analysis and generation are NP-Hard. Koskenniemi and Church (1988) have responded that the complexity problems simply do not arise for natural language, and our experience so far would support that. The NP-Hardness of two-level analysis and generation would become a serious practical problem if natural languages exhibited more phenomena such as vowel harmony, where there are discontiguous dependencies within a word.

In fact, discontiguous dependencies and discontiguous morphemes are a general weakness in traditional two-level systems. In our analysis of Arabic, which has stems consisting of "interdigitated" roots and patterns, we were able to devise enhanced "Detouring" routines that can follow a lexical path through two lexicons at the same time, one containing roots and the other patterns (Beesley, 1989). The result is still a two-level system, showing that the previous inability to handle Semitic languages in two-level morphology can be tied to implementation rather than basic theory.
Because of the way that continuation classes are defined, two-level theory provides a very powerful way to model contiguous dependencies—those dependencies that flow from one morpheme to the next. However, it is quite possible for dependencies to hold between two morphemes that are potentially separated by other morphemes. Consider the following abstract example, where m1, m2, m3, etc. are morphemes in a word.

\[ \text{root} + m1 + m2 + m3 + m4 + m5 \]

Let us assume that the continuation classes specify that a root can be followed by m1 or m2, m2 can be followed by m3, etc. The problem arises if m1 is an optional morpheme which, if present, requires the presence of an otherwise optional m5 later in the word. Unfortunately, in two-level theory the progression on to m5 is controlled only by m4; there is no obvious way to force the analysis to find the m5 morpheme only if m1 appeared somewhere earlier in the word.

One rather unsatisfactory solution is to duplicate all the lexicons that can appear between the interdependent morphemes.

\[ \begin{align*}
\text{root} & \quad m1 \quad m2 \quad m3 \quad m4 \quad m5 \\
& \quad m2' \quad m3' \quad m4'
\end{align*} \]

In such a system, the continuation classes would be defined in such a way that if m1 were present, the analysis would eventually be forced through m5, perhaps picking up m2, m3 and m4 on the way. If no m1 is present, however, the analysis would progress through a duplicate set of lexicons, m2', m3' and m4', where the continuations classes would allow successful analysis of the word without necessarily passing through m5. Similar problems arise if the presence of one morpheme precludes the appearance of another morpheme somewhere in the same word.

Such preclusion was noted in Arabic. Indefinite noun endings (nunation endings) and definite noun endings are only optionally, and very infrequently, realized on the surface. The system must therefore compensate during analysis, postulating definite and indefinite noun endings at the lexical level which are not overtly present in the surface orthography. However, the Arabic word for "the," the definite article, appears bound to the beginning of a noun, as in this definite example.
Normally, a noun that is not marked with the definite article could be interpreted as indefinite, with a nunation ending (the indefinite double Damma is indicated as +N in this example).

Not surprisingly, the presence of the definite article at the beginning of the word precludes the possibility of indefinite (and usually unrealized) suffixes like +N at the end. To avoid duplicating all possible intervening lexicons, we implemented a simple feature unification system (along the lines of Arister, 1987) wherein any morpheme can set certain features for the entire word. For example, the 'al- morpheme can set the feature article=definite, which will clash with the article=indefinite feature set by all the indefinite suffixes. Such feature unification, carried out morpheme by morpheme, serves as an additional kind of filtering during analysis.

For Aymara, we found that the presence of one morpheme could require the presence of another morpheme later in the word. In the following example, the presence of +si with a progressive reading requires the later -ca with a progressive reading.

Such interdependencies were enforced by allowing the morpheme +si to set a feature requirement, such as ca=progressive, a feature which is set only by the required morpheme. Before an analysis path through the lexicons is allowed to succeed, the system now checks to make sure that any and all feature requirements are satisfied by the word's total feature set.

CONCLUSION

ALPNET now has a program which performs morphological analysis and rough word-for-word translation of a significant fragment of Aymara. The morphology rules, presently numbering four, are perhaps complete, and the system is limited mainly by a small root lexicon. We now
believe that the suffix lexicons, containing about 200 separate entries, are nearly complete. We can improve the system by expanding and correcting the dictionaries, continuation classes, feature settings and feature requirements. The Aymara analysis is written using a general morphological analysis tool based on Koskenniemi’s Two-Level Morphology. It is anticipated that we will use the same morphology tool for other languages as the need arises.

The current tool is written in Common LISP and runs on a Macintosh SE with 2.5M bytes of memory. The LISP code itself is rather small, perhaps 25K bytes; the expanded memory is necessary to run the LISP interpreter. We are rewriting the tool in C for porting to IBM mainframes and perhaps other platforms.
Notes


Syntactic Complexity in Virgil’s *Aeneid*

Russell Bateman
Computational Linguist
ALPNET, Inc.
(formerly Automated Language Processing Systems, Inc.)
Provo, Utah

One is struck in the reading of Latin poetry by the sheer syntactic diversity of expression with which the author hopes to control what the reader is feeling. The use of some if not most rhetorical devices and syntactic liberty-taking may have an impact on the reader’s immediate comprehension. In the case of Virgil’s *Aeneid*, it has been posited that a passage of emotional content will often be fraught with syntactic difficulty or conversely that a passage containing difficult syntax will betray emotional content. Challenged by this suggestion, I set out to determine whether there might be some substance to it and to see if I could evolve a simple computational method of identifying syntactic complexity.

**What Is Syntactic Complexity?**

After the previously mentioned suggestion, I become more than vaguely aware as I read that some passages of the *Aeneid* posed more problems than others in comprehension. I began to contemplate different syntactic phenomena and how they might relate to this issue. My initial list included:

- Sentence length
- Ratio of different syntactic categories to the total count of words
- Ratio of modifiers to substantives
- Hypotactic content (relative, subordinate constructions, indirect discourse)
- Parataxis, particularly asyndeton\(^1\)
- Subject, verb and object placement (SVO, SOV, etc.)
- Conservation of the verb *sum*
- Abundance or lack of conditional constructions
- Rhetorical devices\(^2\) which might tend to complicate the text or to render it less readable such as anacoluthon, apostrophe, breviloquenta, chiasmus\(^3\), ellipsis, hyperbaton, synchysis and others which might make it more readable like anaphora and pleonasm.

Given the daunting nature of the task and these virtually limitless parameters, I eliminated a number of items which seemed too great to accomplish. Too, I wanted to succeed in establishing a view of syntactic difficulty which would be based on simple, easily understandable and demonstrable concepts.

---

\(^1\) Or the lack of conjunctions.

\(^2\) An exhaustive list of these, including definitions and examples, may be found in Pharr, Clyde, *Vergil’s Aeneid, Books I-VI*, Grammatical Appendix, items 411—447 (1964, D.C. Heath and Company).

\(^3\) And other liberties with word order and collation.
I chose therefore: the storm stirred up by Juno and Aeolus in the beginning of Book I, Sychaeus' death at the hand of Dido's brother Pygmalion, an excerpt from Sinon in Book II as he weaves his story of his relation to Palamedes, the gobbling-up of Laocoon and his two sons by the sea serpents, the episode where Aeneas returns to Troy in search of his wife Creüsa whom he forgot in his haste to flee the city, and in Book IV, one of Dido's interminable and delightfully contorted soliloquies, the one where she berates Aeneas for thinking of leaving her. Together, these passages represent some 120 lines or about 800 words of syntactically complete text.

Additionally, I chose three passages of at least 100 lines each\(^{11}\) to serve as control to this experiment. Except to exclude the previous six passages, I placed no criterion on their choosing other than the same familiarity of the original subject passages. I took the first from Book II:234-335, which describes the opening of the walls of Troy to the horse, the partying, the city going to sleep, Hector's apparition to Aeneas and the beginning of the sack. The second was chosen from Book IV:522-627, which treats the night before Aeneas' departure from Carthage including another example of Dido's spectacular soliloquies where she tries to reason her way out of fate,\(^{12}\) the visit to the sleeping Aeneas by Mercury and yet another voyage through the raging woman's mind. Last, I chose from Book VI:212-312, a passage dealing with Misenus' proper funeral, the nocturnal sacrifices and the plunging of Aeneas and the Sibyl into Hades and on to their encounter with the ferryman Charon.

**Conclusions**

Following are the statistical results which I will attempt to interpret in conclusion. Here, "parataxis" means nothing more than a count of coordinating conjunctions and "hypotaxis" a count of relative pronouns and subordinating conjunctions.

<table>
<thead>
<tr>
<th></th>
<th>Aeneid as a whole</th>
<th>Control passages</th>
<th>Subject passages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>64215</td>
<td>2048</td>
<td>804</td>
</tr>
<tr>
<td>per full stop</td>
<td>19.01</td>
<td>14.84</td>
<td>26.51</td>
</tr>
<tr>
<td>per half stop</td>
<td>13.27</td>
<td>11.44</td>
<td>15.02</td>
</tr>
<tr>
<td>Parataxis, per sentence</td>
<td>2.05</td>
<td>1.71</td>
<td>2.47</td>
</tr>
<tr>
<td>per 100 words</td>
<td>10.76</td>
<td>11.52</td>
<td>10.10</td>
</tr>
<tr>
<td>Hypotaxis, per sentence</td>
<td>0.37</td>
<td>0.27</td>
<td>0.53</td>
</tr>
<tr>
<td>per 100 words</td>
<td>1.93</td>
<td>1.81</td>
<td>1.99</td>
</tr>
<tr>
<td>Substantives per 100 words</td>
<td>—</td>
<td>32.47</td>
<td>33.67</td>
</tr>
<tr>
<td>Modifiers per 100 words</td>
<td>—</td>
<td>20.31</td>
<td>23.85</td>
</tr>
<tr>
<td>Modifiers/Substantives</td>
<td>—</td>
<td>0.63</td>
<td>0.64</td>
</tr>
</tbody>
</table>

The simplest and most easy phenomenon to understand is the number of words per sentence (a and b) and indeed the results show that sentence length may be the most identifiable feature of complex syntax. Due to the greater length of the sentences, one immediately suspects that Virgil will be employing all manner of rhetorical devices and

---

\(^{11}\) For a total of 2048 words.

\(^{12}\) This may have been better off in the set of experimental passages or excluded from the control set.
word relationships which will contribute to a naturally longer—and as we’ll see—more syntactically complex sentence.

That a greater number of coordinating conjunctions (e) occurred in the subject passages did not surprise me given that the sentences tend to be longer. I took pleasure in noting that on a per word basis (d), it was actually somewhat lower—confirming my suspicion of asyndeton or other some other reason for the omission of conjunctions. Sometimes these are replaced by editors with semicolons, but in spite of this and because of my “half stop” counts, the statistics show that even this device is not completely responsible for the reduced number. Clearly, the passages are heavier for lack of coordinating conjunctions.

I suspect that homography with interrogatives is at fault for the questionable figures for subordinate construction (e and f). It appears that there tends to be a bit more of this in the emotional passages—a more sophisticated syntactic expression. I regard the control passage numbers as more accurate. Perhaps an adjustment to the software to avoid counting words found in initial position would yield more correct and contrasting results. There was little interrogation though a fair number of relative connectives in the subject passages.

The ever-so-slightly higher ratio of substantives and modifiers (g and h) to other categories was predictable and equates directly with sentence length (a and b). For me however, the ambivalence of the ratio (i) yields credibility to my experiment because initial observations had led me to hope for a clear divergence here. That this was not ultimately born out caused me great though temporary concern for I feared that the loss of this “flagship of my hypothesis” would invalidate the entire undertaking and leave me with nothing conclusive to report. Moreover, those counts took the lion’s share of my time since the computer could not perform them. In fact, this caused me to look all the more deeply into the figures I did obtain and I drew conclusions which now seem entirely more reasonable and likely.

Although I have not hit upon an interpretation which will allow me to enumerate unequivocally the elements of Virgil’s syntactic complexity, I see vivid symptoms from this exercise. It now seems clear that the answer lurks in the author’s heavier grammatical relationships forced by a lack of conjunction or rather, that his grammatical device precluded the use of conjunctions in much the same way we in English and French add subordination and otherwise lengthen our sentences in situations where we want to raise our register of speech. In contrast, an interesting question spawned by this conclusion is why we do just the opposite from Virgil when we become emotional; that is, why do we reduce our speech to a rush of short and highly paratactic emissions?

13 I do not propose however nor do I presume to question the current punctuation in the framework of this undertaking.

14 Except for words, the “half stop” statistics are not reported in this condensation of the final numbers. Please refer to the appendix for this.

15 C’est-à-dire, that the numbers for the Aeneid as a whole and for the control passages should be much closer. My suspicion is that somewhere there is an especially great number of interrogatives contributing to such a high ratio for the work as a whole.

16 Indeed, I reported in a short essay in December 1988 covering initial work in Books I and II of the Aeneid that one of the symptoms of complexity would be a higher ratio of adjectives to nouns in the subject passages over the control text.
Methodology

A preliminary process was drawn up and applied experimentally to a small sampling of Books I and II. Out of these experiments quickly emerged encouraging results—confirming that there might indeed be something to the complexity hypothesis. Imposing certain limits and refining some of the steps, I arrived at an easily implementable methodology which consisted mostly of maintaining a set of counters relating to the following aspects:

- **Substantive count.** This count was increased each time a noun was encountered. Excluded were those nouns in the genitive case (voir plus bas). When an adjective was found to be used substantively, it figured in the count also. All pronouns were ignored.

- **Modifier count.** Each time any adjective or noun in the genitive case occurred, this count was increased. Included were pronouns when used as determiners, e.g.: *illo Hectore.*

- **Hypotaxis.** This count was simply incremented each time a subordinating conjunction or relative pronoun was found.

- **Parataxis.** Each coordinating conjunction was counted as it was encountered (*-que or -ue* at the end of words were not forgotten).

In accordance with the foregoing methods and as the subscripts indicate, consider the following three excerpts as examples in applying the methodology:

```
O terque₄ quaterque₄ beati₁
quis₃ ante ora₁ patrum₂ Troiae₂ sub mënibus₁ altis₂...

raptatus₂ bigis₁ ut quondam, ...

uel Danaum₂ Phrygios₂ iaculatus₂ puppibus₁ ignis₁!
```

I thought it might be useful for the purposes of the word-to-sentence ratios to break sentences not only in the traditional places marked by the period and the question and exclamation marks (called *full stops* in some terminologies), but also at what I will call, for want of a better term, *half stops* including the colon and semicolon which always fall at complete syntactic pauses. The results of this appeared in the larger set of reported statistics and confirmed what was already born out through the normal breaks.

To speed up the process and enable it to include as much of the *Aeneid’s* text as possible, the implementation of the counting process was divided between that which could be done by the computer and that which required biological intelligence. Of course, it was necessary that the software require only a minimal amount of time spent in design and implementation of its programs. An eye was also to be kept toward moving more and

---

4 A nominalized adjective.

5 This may have been a mistake though it ultimately eliminated some subjective decision making in the counting process.

6 This encourages the thought that a mistake was made previously in leaving out pronouns although few instances of this actually occurred.

7 This was occasionally difficult—and close attention was necessarily paid to this issue.
more of the computational process from intelligent activity to the machine as increased understanding of the problem permitted.

The machine first began counting only words and sentences. All ratios could of course be produced quickly and accurately on the computer. Then, by brute force, coordinating and subordinating conjunctions as well as relative pronouns were counted this way since they all exist in lexically finite sets and since due to time constraints—as mentioned earlier—I gave up much of what I wanted to include in counting instances of hypotaxis, particularly indirect discourse. In fact, hypotaxis became too big a concept considering what was ultimately placed under this heading. It remains however a useful alias for two other simple but very long terms: subordinating conjunction and relative pronoun. Similarly, though more accurately, parataxis ended up only including a count of coordinating conjunctions.

For lack of a dictionary as well as a morphological reduction and look-up capability, there remained the task of counting substantives and modifiers. This seems simple at first glance: simply count nouns and adjectives. But both notions become more complex considering the vast verbal system of Latin. Still, a simple approach needed to be found in order to complete the study and be able to interpret the results.

A number of practical difficulties plagued me as I made the counts. It was often hard in spite of my familiarity with the text, to determine when to count a present or past participle as a modifier and when to leave it out as a verb. By convention, one is encouraged to construe Latin participles as clauses, but since such a construction might yield a clause simply containing a copula (when not deponent of course in the case of the past participle), it still nevertheless passes as a modifier. On the other hand, similar clausal phenomena, among them relative clauses of characteristic, etc., were not counted as modifiers since this violated the criterion of simplicity. In view of this, there was a tendency to leave out many past participles as well. Ultimately, I made the choice based on how “adjectival” I felt the form was at the moment of choice and left it at that.

Other problems surfaced in the computer program I wrote to make the easy counts and later the subordinate construction counts. Some errors were almost certainly produced by over-detection of quis and its composites in the case where it has supplanted aliquis. Interrogative pronouns when homographic with relatives and subordinating conjunctions caused the program to err a bit in favor of hypotaxis. All errors committed by the software were done so uniformly across the entire Aeneid however.

**Choice of Passages**

I selected six passages of greatly differing length in the most subjective manner possible, namely when I felt that I had had a difficult time deciphering the passage or when I felt that it had good potential as an emotional passage. While there is perhaps some flaw in the latter with respect to scientific method, I was impatient to see whether statistical analysis of the passage supported the conclusions I hoped to draw.

---

8 The main program and all supporting software was implemented in the C programming language under SunOS (a derivative of Berkeley UNIX) on a Sun Microsystems 3/50™ using a corrected version of the IBYCUS text.

9 In this case, simple boolean searching.

10 I used no passages which I hadn’t read at least two or three separate times nor with which I felt any insecurity.
Complete statistical output for the *Aeneid* as a whole and for each book and the passages examined in this paper, as well as the C source code listings which are too long to reproduce here, are available at no charge from the author.

Russell Bateman
742 West 400 North
Provo, Utah 84601
(801) 375-8733
Natural Language Processing Applied to a Data Base of Engineering Design Documents

Larry E. Wood, Randy L. Orr, and Alan R. Parkinson
Brigham Young University

Abstract

The paper describes an application of Roger Schank's approach to natural language processing by computer, called Conceptual Analysis, which focuses on the underlying conceptual representation of an event. By definition, an event involves an action, an actor, an object, and a direction to which the action is oriented, plus other aspects that might be relevant (e.g., time, place, etc.). The particular problem to which Conceptual Analysis has been applied is a database of engineering design documents. A prototype system was developed to analyze functional descriptions of design documents, allowing it to recognize that components of different designs perform similar functions.

Introduction

This research describes a prototype intelligent system for the storage and retrieval of function-oriented descriptions of engineering design documents. While functional concepts have potential utility to the design process, they pose some difficulty for storage and retrieval systems because a functional description requires a phrase or sentence and the same concept can be described in a variety of ways. For example, consider the phrases "increase speed" and "amplify rate of motion." These phrases are composed of different words but they imply the same underlying concept of "varying" the magnitude of an attribute (velocity) of an object.

What is needed to store and retrieve a design function description is a natural text processing scheme that can "understand" the underlying concept that a specific function represents, so that designs performing similar functions can be recognized. Such a scheme, referred to as Conceptual Analysis (Cullingford, 1986) was used to develop a prototype system for storage and retrieval of functional descriptions of design documents. The system is called FORRDD for Function-Oriented Representation and Retrieval of Design Documents.

Conceptual Analysis

Conceptual Analysis is based on Conceptual Dependency theory developed by Roger Schank (e.g., Schank, 1973), a method of natural language processing based primarily on semantics rather than syntax. Its foundation is the notion of an underlying conceptual representation. Schank proposed three types of elemental concepts: nominals, actions and modifiers. Such things as a man, a duck, a book and a pen are nominals. An action is what a an animate object does to another object. Some examples are
"John hit Bill", where "hit" is the action or "Mary took the book", where "took" is the action. Modifiers are such things as "red" in the phrase "the red ball" or "slowly" in the sentence "He approached slowly".

In order to capture the underlying conceptual representation of an event, Conceptual Dependency focuses on the action being performed. Schank claims that all words describing actions can be mapped onto a relatively few primitives. Each primitive has its own concept frame containing "slots" that can only be filled by specific nominals or modifiers. It assumes that a statement refers to some type of an event having the following aspects (at least):

- an ACTOR
- an ACTION performed by that actor
- an OBJECT that the action is performed upon
- a DIRECTION in which that action is oriented

By definition, an event contains all of the above elements and if not explicitly mentioned in the sentence, a hearer/reader is caused to infer actors, objects etc. One of the key functions of the system is slot filling. Initially empty slots are filled by items obtained from an analysis of the sentence or from "world knowledge" previously entered into a knowledge base.

One of the primitives proposed by Schank, called ATRANS, refers to actions related to transfer of possession. It encompasses such verbs as "give", "take", "buy", "sell", "rob", and "borrow". An example of a sentence involving the primitive ATRANS would be "John gave Mary a book." A conceptual analysis program based on Conceptual Dependency analyzes such a sentence from left to right, searching for elements such as the primitive action, its actor, an object transferred, and the direction of transfer. The resulting analysis fills in the initially empty conceptual frame producing a meaning structure such as the following:

```
ATRANS  Actor (Person Name (John))
         Object (book)
         To (Person Name (Mary))
         From (Person Name (John))
```

As indicated earlier, the FORRDD system actually uses a conceptual analysis system based on work by Richard Cullingford (Cullingford, 1986), which is an outgrowth of Conceptual Dependency. In addition to characteristics of Conceptual Dependency discussed above, Cullingford's system imposes a basic classification scheme on common sense knowledge, which he refers to as Eclectic Representation for Knowledge Structures (ERKS).

The ERKS system consists of entities, acts, and relationships which are subdivided into a number of primitive types of conceptual structures. These types are organized into a simple inheritance hierarchy (sometimes called a ISA hierarchy, because items share or inherit attributes from items above them in the hierarchy). Thus, when we say "a person ISA animate object," what we mean is that certain things true of animate objects are true of people.

152
Imposing such an organization of primitive classes provides a standard, useful means of controlling the inference processes that are concerned with reasoning about primitive classes. Problem-solving processes trying to determine, for example, whether a book is a suitable "object" of a giving action, can test if "book" is a member of the class of movable or transferrable objects. Once primitive classes have been developed for a particular domain of analysis, then words that will appear in the sentences and/or phrases to be analyzed must be defined for the system. The word definitions contain two types of information: (1) associations between the word and specific conceptual structures and (2) expectations or predictions about the other concepts in the context surrounding the word as it is used in a sentence.

Functional Design Concepts

As was discussed earlier, an important aspect of Conceptual Analysis is the development of a set of primitive concepts which capture the underlying meaning of a variety of similar language descriptions. For storage and retrieval of design documents, the FORRDD system was inspired by the hypothesis that functional characteristics or concepts would be useful because they are independent of particular implementations (see Pahl & Beitz, 1984). Thus, in a company that manufactures both hand tools and small appliances, a designer might not recognize close similarities in terms of physical characteristics. However, there likely could be components in a small appliance that perform functions similar to those in a hand tool, and these would be made available to a designer through a retrieval system based on functional concepts.

Pahl and Beitz (1984) propose that the overall function of a design expresses the relationship between system inputs and system outputs, which can be reduced to three forms of matter in the case of electro-mechanical designs: energy, material and signals. The functional relationships between inputs and outputs can then be conceptualized as the conversion in various ways of these three forms of matter.

Another important aspect of the work of Pahl and Beitz (1984) is their notion of establishing function structures during the design process. This consists of dividing the overall function of a design into sub-functions recursively until a point is reached where a sub-function cannot be sub-divided further and still remain generally applicable. Pahl and Beitz refer to functions at this level as generally valid functions. They propose the following set of generally valid functions: CHANGE, VARY, CONNECT, CHANNEL AND STORE.

CHANGE involves the transformation of one type of matter to another. VARY applies to changes in the magnitude of a particular type of matter. CONNECT refers to conversions resulting in differing numbers of inputs and outputs. CHANNEL designates changes in location or place. STORE refers to time-dependent functions. The mapping of design functions to generally valid functions will be illustrated with some examples from an electric can opener. The overall function "Remove the lid from a can" could map to a CONNECT with one input (the can) and two outputs (the can and the lid). The function of a gearbox assembly to
"Transfer torque from a worm gear to a drive shaft," could map to a CHANNEL. Finally, an electric motor with a function to "Transform electrical energy into mechanical energy" could map to a CHANGE.

Conceptual Analysis of Functional Design Concepts

As was discussed earlier, an important aspect of Conceptual Analysis is the development of a set of primitive concepts which capture the underlying meaning of a variety of similar language descriptions. In the FORRDD system, the five generally valid functions proposed by Pahl and Beitz (1984) were used to fill that role. First a conceptual structure was developed for each generally valid function, establishing the various slots that would be required to accomplish it's task. For example, because CHANNEL pertains to changing the location of some type of matter, it was necessary to provide slots for the item to be relocated (Channellee), the input location (InLoc), and the output location (OutLoc). An additional slot (Actor) was also added to accommodate the name of an assembly that performed the function in question. This results in a conceptual structure such as the following:

(Channel Actor ( )
Channellee ( )
InLoc ( )
OutLoc ( )

Following the development of conceptual structures, a domain of designs was chosen for developing a prototype system. In this case the domain of small electrical appliances was chosen, consisting of the following items: a can opener, a hand mixer, a knife sharpener, a toaster, and a coffee maker. For each of these, sub-assemblies were described including such components as gear boxes, electric motors, shafts, gears, filters, heating elements, etc.

A set of functional descriptions of the items, assemblies, sub-assemblies, and components in the domain were then produced, thereby establishing the entities which the system would have to account for and the vocabulary that would be required to analyze the descriptions. Next, an ERKS hierarchy of the types of objects (e.g., gears), and entities (e.g., types of matter, energy, and material) was constructed. Finally, definitions for all the words contained in the set of functional descriptions was developed. The verbs are listed in table 1 and the nouns and adjectives are shown in table 2. A sample description, "A drive shaft transfers torque from a worm gear to a drive gear", will be used to illustrate this process in a simplified form. First a definition was derived for "drive shaft" by associating it
Table 1 -- Verbs (indented words are synonyms) currently defined in the FORRDD system

<table>
<thead>
<tr>
<th>transfer</th>
<th>support</th>
<th>turn</th>
<th>transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>transpose</td>
<td>connect</td>
<td>move</td>
<td>regulate</td>
</tr>
<tr>
<td>direct</td>
<td>mount</td>
<td>lower</td>
<td>control</td>
</tr>
<tr>
<td>shift</td>
<td>mix</td>
<td>lift</td>
<td>vary</td>
</tr>
<tr>
<td>convey</td>
<td>pierce</td>
<td>raise</td>
<td>decrease</td>
</tr>
<tr>
<td>transmit</td>
<td>sheer</td>
<td>remove</td>
<td>reduce</td>
</tr>
<tr>
<td>channel</td>
<td>penetrate</td>
<td>release</td>
<td>increase</td>
</tr>
<tr>
<td>conduct</td>
<td>cut</td>
<td>store</td>
<td>amplify</td>
</tr>
<tr>
<td>guide</td>
<td>hold</td>
<td>generate</td>
<td></td>
</tr>
<tr>
<td>deliver</td>
<td>clamp</td>
<td>create</td>
<td></td>
</tr>
<tr>
<td>receive</td>
<td>secure</td>
<td>indicate</td>
<td></td>
</tr>
<tr>
<td>locate</td>
<td>rotate</td>
<td>convert</td>
<td></td>
</tr>
</tbody>
</table>

with a simple conceptual structure that specified it to be an assembly which is a type of material, which is a type of matter, which is a type of physical object. When the system encounters the term "drive shaft" the following conceptual structure is simply added to a concept list: ( Assembly Type (Driveshaft)).

The word "transfer" was mapped to the primitive CHANNELL as described above. When "transfer" is read by the system, the first thing that happens is that an empty transfer concept is added to the list as follows:

( Channel Actor
   Channellee
   InLoc
   OutLoc )

As might be expected, the action words "do the most work" and therefore have the most complex definitions. In addition to associating "transfer" with a CHANNELL structure and adding it to the concept list, its definition also must activate expectations (or requests as they are called) for concepts that can fill the four slots in the CHANNEL structure. The request to fill the Actor slot looks for something already on the concept list that could be an assembly, will find the concept created for "drive shaft", and will fill the Actor slot with it. The other requests must wait for additional concepts to be created as the remaining words are processed.

The definitions for "torque", "worm gear", and "drive gear" are similar to the one for "drive shaft" in that each is associated with a simple conceptual structure. "Torque" is associated with a type of motion (rotation) which is a type of energy which is a type of matter. The request to fill the Channellee slot searches for a type of energy, and uses the concept added by "torque" to do so. "Worm gear" and "drive gear" are both types of assemblies as was "drive shaft". The requests for the InLoc and OutLoc slots search for types of assemblies following the
Table 2. Nouns and Adjectives (indented words indicate sub-types) currently defined in the FORRDD system

<table>
<thead>
<tr>
<th>signal</th>
<th>rotational</th>
<th>motor</th>
<th>filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>fluid</td>
<td>torque</td>
<td>filter</td>
<td>housing</td>
</tr>
<tr>
<td>water</td>
<td>translation</td>
<td>worm gear</td>
<td>motor</td>
</tr>
<tr>
<td>coffee</td>
<td>linear</td>
<td>worm gear</td>
<td>flow</td>
</tr>
<tr>
<td>matter</td>
<td>oscillation</td>
<td>shaft</td>
<td>regulator</td>
</tr>
<tr>
<td>input</td>
<td>counter-rotation</td>
<td>drive gear</td>
<td>temperature</td>
</tr>
<tr>
<td>output</td>
<td>rotation</td>
<td>supporter</td>
<td>sensor</td>
</tr>
<tr>
<td>energy</td>
<td>mechanical</td>
<td>shaft</td>
<td>sub-assembly</td>
</tr>
<tr>
<td>power</td>
<td>friction</td>
<td>spring</td>
<td>unit</td>
</tr>
<tr>
<td>force</td>
<td>material</td>
<td>drive shaft</td>
<td>rpm</td>
</tr>
<tr>
<td>electricity</td>
<td>can</td>
<td>bushing</td>
<td>mph</td>
</tr>
<tr>
<td>current</td>
<td>lid</td>
<td>knife</td>
<td>ft/s</td>
</tr>
<tr>
<td>electrical</td>
<td>food</td>
<td>knife blade</td>
<td>lb</td>
</tr>
<tr>
<td>potential</td>
<td>lubrication</td>
<td>holder</td>
<td>hz</td>
</tr>
<tr>
<td>kinetic</td>
<td>appliance</td>
<td>beater</td>
<td>inch oz</td>
</tr>
<tr>
<td>heat</td>
<td>can opener</td>
<td>heat element</td>
<td>ft lb</td>
</tr>
<tr>
<td>temperature</td>
<td>knife</td>
<td>assembly</td>
<td>location</td>
</tr>
<tr>
<td>hydro</td>
<td>sharpener</td>
<td>gearbox</td>
<td>direction</td>
</tr>
<tr>
<td>flowrate</td>
<td>coffee maker</td>
<td>can opener</td>
<td>orthogonal</td>
</tr>
<tr>
<td>motion</td>
<td>mixer</td>
<td>cutter</td>
<td>plane</td>
</tr>
<tr>
<td>speed</td>
<td>toaster</td>
<td>electric knife</td>
<td></td>
</tr>
<tr>
<td>rotation-</td>
<td>component</td>
<td>toaster</td>
<td></td>
</tr>
<tr>
<td>circular</td>
<td>gear</td>
<td>lifter</td>
<td></td>
</tr>
</tbody>
</table>

prepositions "from" and "to", respectively. Thus when the two prepositional phrases are encountered, the concepts are added to the concept list and promptly used to fill the InLoc and OutLoc slots. Thus, the analysis results in the following meaning structure:

(Channel Actor (Component Type (Driveshaft))

Channel (Motion Type (Rotation)

InLoc (Component Type (Wormgear))

OutLoc (Component Type (Drivegear)))

As an indication of the versatility of the Conceptual Analysis system, the definitions and ERKS types discussed above would be adequate for the following collection of phrases and sentences:

(a) A drive shaft transfers torque from a worm gear to a drive gear.
(b) A drive shaft transfers torque
(c) A drive shaft transfers.
(d) Torque is transferred by a drive shaft to a drive gear from a worm gear
(e) Torque is transferred from a worm gear to a drive gear
(f) Torque is transferred.
(g) Transfer.
Once all the ERKS hierarchy and all the necessary word definitions were developed, the storage and retrieval system was developed. That aspect of the system will now be described.

Storage and Retrieval System Description

Storage of functional design concepts involves two basic procedures. The first is the creation of a full text file of a design document, including specifications and any information about assemblies and sub-assemblies that might be desired. The second is the entering of functional descriptions of the assemblies and sub-assemblies described in the design documentation into a database of descriptions and linking them to the design document file. In the current implementation, the document file can be created by typing the information directly the system or by a file transfer from another system.

Entering functional descriptions of assemblies into the database is accomplished by following system prompts. First, the system prompts for the name of the document file to which the functional descriptions apply. This would typically be the name of the object (appliance, sub-assembly, or component) which the document describes. Next, the system prompts for the functional description of an assembly such as the example discussed earlier, “The drive shaft transfers torque from the worm gear to the drive gear.” The system then analyzes the description, produces a meaning structure of the type discussed earlier, adds the name of the file to a “PartOf” slot in the structure, and adds the structure to the database of meaning structures. Additional descriptions are added as needed in this manner until all assemblies and sub-assemblies in the document have been entered.

Retrieval of document information is then accomplished in a similar manner to entry, except that the user is prompted for a general functional description not including the name of an assembly, e.g., “Transfers torque from the worm gear to the drive gear.” The system analyzes the description and searches the database for one that matches. If a match occurs, the system uses the assembly name (stored in the Actor slot) and the file name (stored in the PartOf slot) and displays a message such as, “That function is performed by a drive shaft which is part of a Can Opener.” The user can then review the relevant file containing the design document.

Currently, the matching algorithm will produce a match to a meaning structure (based on the same primitive function) that is at the same level of specificity or more specific than the query. In the example given, the user could omit either “from the worm gear”, “to the drive gear”, or both. The less specific the query, the more likely it will result in a match, as with most database queries. Results of queries varying from general to specific are shown in Table 3.
Table 3 -- Retrieval Examples

1. A query consisting only of the word "transfer", will result in 25 items such as: (a) housings for each of the appliances supporting sub-assemblies (transfer of supporting energy), (b) gears transferring mechanical energy from a source to a destination, (c) a flow regulator transferring signals to regulate the flow, and (d) heating element transferring heat to water or to bread.

2. A query consisting of the phrase "transfer energy", results in 15 items, now restricted to housings, gears, drives shafts, etc.

3. A query consisting of the phrase "transfer torque" (a particular type of energy), results in 8 items, now restricted to drive shafts and gears.

4. If you ask for things that "transfer torque of 45 in. oz.", results in 1 item, a drive shaft that is part of a can opener.

Conclusions

The objective of this research was to develop a prototype system that applies natural language processing to design classification for the purpose of storing and retrieving designs by their functional concepts. The prototype has thus far demonstrated the feasibility of a retrieval system based on functional design descriptions. However, it has yet to be demonstrated that such descriptions will be useful to designers. According to French (1988), functional design has not been studied extensively, and there is not yet much evidence that engineers think explicitly in terms of function during the design process.

While the development of FORRDD is centered around functional design concepts, there are certainly other design concepts which might be useful for retrieval purposes (e.g., physical specifications, maintainability, reliability). Further work in this area is needed. Indeed, Wasserman and Lebowitz (1983) have reported a system, called RESEARCHER, which uses conceptual dependency to analyze patent descriptions in terms of physical characteristics and relationships.

In addition to those discussed, there are several considerations that simply need evaluation by testing the system and the principles behind it on designers in a realistic setting. We are exploring various means by which such an evaluation might be accomplished.
References


Notes

1The research described here is based on work performed by the second as part of the requirements of an M.S. degree in engineering technology.
At last October's annual conference of the American Translator's Association in Seattle, Charles Teubner of the Princeton Technical Translation Center in New Jersey explained to a very interested audience how he uses Systran, one of the oldest and largest machine translation companies. His source documents are written in Mass-11, which is a word processing system especially suited for representing scientific formulas and equations. He said that in order to send his documents through the Systran system, he first has to strip out the formatting commands by hand, and then after receiving his translations back from Systran, he must put the formatting commands back in manually as well.

Based on my own experience with computer translation, I found his situation to be quite indicative of the state of much commercial machine translation today. There are several interesting points which can be drawn from his story.

First, machine translation (MT) is being used today in a commercial setting. This proves that despite the numerous linguistic problems and difficulties, machine translation can be a viable, productive tool. And although I got the impression that Mr. Teubner does not use MT on large quantities of material, there are many other businesses and organizations today which do use the products of the various MT companies successfully on very high volumes of text.

Second, source texts which are used in commercial machine translation systems are often created in machine-readable format. Typical source texts for MT include user manuals for everything from computers to tractors, and nowadays these are almost all created on word processors. This fact has helped to alleviate the considerable problems associated with first converting source text into machine-readable form in order to use it in machine translation.

The third point which can be drawn from this story, and which I wish to elaborate on in this paper, is that even when source texts are created in machine-readable form to begin with, there are still problems involved in getting the texts to work for machine translation, or at least to work well. In other words, it is not always easy to get machine-readable source text into a format that can be used by a machine translation system.
Mr. Teubner is obviously keen enough on the idea of machine translation to take the time to adapt his texts by manually removing and replacing formatting commands in order to use an MT system. But this manual method is hardly an economical way to do machine translation, and for large volumes of text it becomes far too time consuming to be feasible at all.

Systran might be criticized for failing to provide some sort of automated solution to this problem, but they cannot be singled out for blame. It has been my experience that none of the other commercial MT companies have done much better. It has only been within the last year or two that the company for which I work, ALPNET, has seriously addressed the issue.

I think MT companies have been slow to respond to this need because they have failed to appreciate the importance and extent of the problem. Systran and most of the other MT companies came into being before the virtual explosion in the number of different word processing systems over the past several years, and consequently their systems were not originally designed to deal with a variety of word processing formats. As the need has arisen, the various MT companies have underestimated the difficulty in adapting the various word processing formats for use by machine translation systems, assuming that it was a relatively simple logistical problem. However, if not dealt with adequately, this problem can destroy the economic viability of a commercial translation system. To paraphrase an old saying: "For want of usable source text, the machine translation system was lost."

What exactly are formatting commands, and why are they so difficult to deal with? Every document, i.e., computer file, that is created with the use of a word processor contains not only the text that the author typed in, but also a whole host of typesetting information, known as markup or formatting commands, which indicate such things as margin width, tab settings, spacing, centering, justification, pitch, font, table of contents and index entries, bolding, italics, and underlining, to name some of the most common ones. Depending on the word processing system, some commands are placed automatically in the file as default values. Others are put in by the author as he or she wishes to modify the format of the document.

These commands are generally character strings, just like any other regular word in the file, but they have special meaning to the word processing system and are generally not visible in the printout of the word processing document. Within the computer file itself, however, these commands can occur virtually anywhere. Some types of commands occur in blocks between paragraphs or at the beginning of the file; others are interspersed with the regular linguistic text of
the file. And depending on the type of word processing system used, it may not be immediately obvious to the typist that some of these commands are being mixed in with the regular text.

The placement of markup commands is illustrated in Appendices 1 and 2. Appendix 1 shows the printout of a text created by the WordPerfect word processing system. Appendix 2 shows the same file with the markup commands visible.

When these markup commands occur within sentences, it wreaks havoc with the output of a machine translation system. The commands are interpreted as "words" in the sentence by the computer, but of course, since these "words" have no grammatical function, they only serve to confuse the computer analysis. The resulting translation is often so garbled, that the human revisor has to retranslate the entire sentence manually, thus negating the advantages of using the computer for translation.

In order to use word processing files at all in machine translation, the markup commands must first be removed from the linguistic text of the source document. And then, if the translation is in any way to reflect the same format as the source document, the markup commands must be put back in the translated text. It is possible, of course, to do all this by hand, but the only really economical method is to do it automatically.

Automatically removing the markup commands can be fairly trivial; many word processors can do this as a standard feature. The difficulty for machine translation comes in automatically putting them back into their corresponding positions in the translation. Because the word order of the translation is often quite different from that of the source, it is difficult even to define exactly what "corresponding position" means.

Let me illustrate the problem by discussing a simple algorithm for dealing with markup commands. Before we remove the commands from the source text, let us mark their position by associating each command with the word next to it. In our algorithm, we will arbitrarily associate each command with the word to its right. Then in the translation, each command is put back in the text to the left of the translation of the word with which it was associated.

This method works well for certain commands, such as "[FtnOpt]" (footnote option) in Appendix 2, whose exact position in the sentence is not critical. The [FtnOpt] command merely specifies the printed format of subsequent footnotes and can appear anywhere in the text before the first footnote whose format is to be specified. In this algorithm, the [FtnOpt] command in Appendix 2 would be
associated with the word "with" and would then appear to the left of the translation of "with" in the target text.

For other types of commands, where exact position in the sentence is more important, the results of this algorithm can be disastrous. Take highlighting commands, for example. These are commands such as "bold", "underline" and "italic", which indicate how a word or a phrase should be highlighted. These commands come in pairs: a begin highlight, and an end highlight command.

Figure 1 shows an example of the underline commands in WordPerfect. Everything between the begin highlight ([U]) and the end highlight ([u]) is to be underlined. Figure 2 shows the results of applying our algorithm to these commands when translating into Spanish. Instead of "arte moderno" being underlined, only "moderno" is.

Fig. 1  He studies [U] modern art [u] in Madrid.

Fig. 2  Estudia arte [U] moderno [u] en Madrid.

We may be tempted to say that this algorithm doesn't work merely because of the arbitrariness of associating each command with the word to the right. It is obvious that end highlight commands stand in a relationship with the word to the left, not to the right. In our example, [u] is really associated with "moderno", not "en".

However, if we refine our algorithm to associate end highlight commands with the word to the left, the results, shown in figure 3, are even worse. We end up with the end highlight command before the begin highlight.

Fig. 3  Estudia arte [u] [U] moderno en Madrid.

Actually, all algorithms based solely on associating formatting commands with adjacent words are insufficient. The best way to handle highlighting commands is to mark each word between a begin and end highlight command pair as belonging to that particular highlighted phrase, and then removing the highlight commands from the source text. In the translation, a begin highlight command is inserted between each pair of words where the left word is not marked for highlighting, but the right word is. The end highlight command is inserted between each word pair where the left word is marked and the right word is not.

This algorithm also works well when a highlighted phrase is split into two or more non-contiguous groups in the translation. Take, for example, the single English word "open" which is highlighted in the sentence in figure 4. When translating into German, this single word is split into two non-contiguous words. See figure 5. In this case, not
only do the original commands have to be replaced, but new commands have to be generated to surround each split-off element.

Fig. 4  [U] Open [u] the window.

Fig. 5  [U] Machen [u] Sie das Fenster [U] auf [u] .

It is even possible to highlight only part of a word, as in figure 6. However, unless the computer somehow maintains a correspondence of morphemes between each source and target word, which would be very difficult, if not impossible, to do, there is not a perfect solution to this problem. Perhaps the best approach in this case would be to replace the end highlight command at the end of the word from which it was taken. Conversely, if a begin highlight command is in the middle of a source word, it could be put back before the beginning of the translation of the word. This would leave the entire word highlighted in the target, but manual post-editing could easily restore the original format.

Fig. 6  [U] under[u]line

There are still other types of markup commands which are problematic. Sometimes the command itself contains text which should be translated. The WordPerfect footnote command, for example, contains the entire text of the footnote, yet in the internal file structure, the footnote text is placed within the footnote marker inside the original sentence. See Appendix 2. In this case, the command must be removed in order to translate the original sentence, but the text within the command still has to be presented to the machine translation system at some point in order to be translated itself.

Other commands, even though not needing to be translated, fulfill some linguistic function and should be left in the sentence when it is translated. IBM Corporation's ISIL markup system ("Information Structure Identification Language"), for example, has commands known as "variables" which stand for nouns, usually product names. At some point in the processing of the document, the variables are replaced by the nouns that they stand for, but during translation (and IBM uses machine translation to translate many of its documents), they are generally still variables. These commands typically function as subjects or objects in the sentence and must be considered as nouns by the machine translation system.

The same sort of command exists in word processing systems which have a "form letter" or "merge" capability. Appendix 3 shows an example of a Microsoft Word form letter. The words and phrases in the letter which are enclosed by guillemets (« ») and begin with lower case letters are special
variables, called "field names", and are linguistically significant. (The other commands, viz. IF, ELSE and ENDIF, are discussed later in this paper.) Form letters also have an accompanying "data document" which contains the "values" for these variables. In this case, the data document consists of a list of customers' "name"s, plus the "product" that each customer is interested in along with the "retail" price. The form letter and the data document are then "merged" when printing; i.e., a different, personalized letter is printed out for each customer in the data document.

In some word processing systems, these linguistically significant commands can stand for more than just nouns. ISIL, for example, also contains a whole set of markup commands which stand for different kinds of punctuation marks, such as colons and quotation marks, and which, for various formatting reasons, are typically used in a document rather than the actual punctuation. A machine translation system, of course, must recognize each different command and treat it as the punctuation mark for which it stands. There may be dozens of different commands that have to be treated individually for each word processing system that is handled, which can make the automatic handling of markup commands extremely complicated.

Finally, let me give one more example of a class of markup commands, which to my knowledge has not yet been encountered enough in commercial machine translation to become a problem, but which does illustrate some of the nastier possibilities which markup commands can present. Some word processing systems have a series of commands which essentially turn the text file into a sort of computer program, and can change the final form of a document quite drastically depending on the value of certain "variables". Appendix 3 illustrates the Microsoft Word, "IF", "ELSE", and "ENDIF" commands. In the form letter shown, the text of the final paragraph reads quite differently depending on the value of the "retail" variable. In a machine translation system which must grammatically parse each sentence, this sort of thing can be quite a problem indeed.

In conclusion, word processing presents an interesting challenge to the machine translation industry. In the past, the problems of handling markup commands have been overlooked or dismissed as fairly simple logistical details. In fact, dealing adequately with markup commands is critical to the economic usefulness of machine translation, and some of the problems involved in dealing with them are not trivial at all.
Sometimes a phrase is underlined to emphasize a point. At other times boldface type is used. Combining highlights lends extra emphasis to what has been written.

Another common practice is to indent entire paragraphs in order to set them off from the rest of the text. In this paragraph, the pitch has also been changed from pica to elite, and the right margin has been justified, which highlights the paragraph even more.

1This is the body of the actual footnote.
Dear «name»,

Thank you for your request concerning our «product». The «product» sells for «retail» plus 5% tax and handling. If you wish to order the «product», please remit a total of: «retail + retail * .05».

We would also be glad to send you our catalog at «IF retail > 10» no extra charge «ELSE» a cost of only $2.00 «ENDIF».

Sincerely,
When the computer age began, the myriads of possibilities for the computer's use in business, government and education were still unfathomable. Expressions such as "computer assisted instruction" and "hypermedia" were used only by scientists and programmers. Nevertheless, in a few short years the numerous capabilities of the computer were manifest. Today, "software," "floppy disk," and "P. C." are a part of the average person's vocabulary. Every day children "boot up" to learn such fundamentals as reading, writing, and arithmetic. Of late, programs teaching foreign language have been added to these instructional materials. This paper will discuss one such foreign language program developed to supplement a beginning Spanish course. CASA, an acronym for Computer Assisted Spanish Applications, was developed at Brigham Young University by Earlene Sudweeks and Barbara Gordon, two undergraduate students in the Spanish program.

The CASA program is developed using HyperCard on the Macintosh computer. HyperCard creates software that teaches, tutors and instructs. It is relatively simple to master the HyperCard language and to use this language to develop professional-looking software. "Stacks" and "cards" form the basis for HyperCard programming. Each screen displayed on the computer terminal represents a "card". These cards are grouped together to form stacks, and one stack or a combination of several stacks constitute a program. The applications in HyperCard use such supporting structures as tools, graphics, and sound.

The tools of Hypercard include such options as "buttons" and "fields." Buttons are used to execute the algorithm of the program. They allow various texts to appear and disappear at will, they "turn the pages" of the HyperCard cards. These features increase the interest of the student as they allow him or her to interact with the program. Fields also play a very important role as a HyperCard application. They function mainly to present text, however, their use is versatile.

HyperCard also contains tools and capabilities to make use of or create various types of graphics. Graphics created with paint tools or scanned images allow the programmer to visually stimulate the user.

Digitized sound is another characteristic of HyperCard. Sound is stored as a resource in a given stack and is activated by using a "play" statement in the script. Sound can be incorporated into the program by a MacRecorder which accepts both prerecorded and directly recorded sounds. Direct recordings are those sounds spoken directly into the MacRecorder where editing can take place immediately. HyperCard has a complete programming language called HyperTalk. The scripts
within the stack, written in HyperTalk, are what give all buttons and fields and other such HyperCard applications their maneuvering capabilities.

The following will more fully explain the characteristics of the aforementioned HyperCard properties as used in the CASA program.

The first stack encountered in CASA is an introduction to the program itself. Automatically, music is played while an invisible paint brush tool paints an illustration of a house (casa in Spanish) and places "Computer Assisted Spanish Applications", along with the acronym CASA, inside the drawing. This opening card, with the drawing and sound, serves to capture the attention of the user as he immediately realizes that this will not be a monotonous tour through endless cards of explanations and detailed descriptions.

Following this brief introduction the user is given an overview of the Spanish alphabet and of the basic pronunciation rules, all of which include sound resources to provide the user with the correct pronunciation of each. A short quiz follows to test the student's comprehension level.

Once the user has mastered the alphabet and pronunciation, he is encouraged to continue on to the grammar portion of the stack. Keeping in mind that this program is geared toward the beginning-Spanish student, the grammar is presented in a simplistic and understandable manner. First, the pronouns with the English and Spanish equivalents are taught. Then a few regular verbs are presented in full conjugation in the three most common tenses: present, past, and future. Through a series of well-constructed cards, the user has the option of reviewing as often as necessary the presented material both in Spanish and in English before continuing with the program.

The program then continues with a presentation of common dialogues. The objective of this section is to enhance the listening-comprehension skills of the student by presenting three typical conversations that would commonly occur in a park. Both the written and audio dialogues are given.

Once again the format changes and the student is given an opportunity to test his reading skills and to learn additional vocabulary. A short children's story, Hansel and Gretel, is displayed in book-like form complete with open pages and illustrations. Each page contains text which has hidden buttons that allow the user to "click on" unfamiliar words to see their English equivalent. Sentence by sentence, the reader is given the opportunity to not only read the text but to hear it spoken. By this he can add to his understanding of spoken Spanish as well as written Spanish. Following this portion, the user is once again presented with questions to test his level of comprehension.

Another feature was added to CASA that promotes its usefulness. Realizing that many of the beginning-Spanish students will one day face foreign airports, non-English speaking taxi drivers and hotel receptionists, CASA takes the student on a trip to a spanish-speaking country. Simulations and dialogues allow the user to master the vocabulary necessary for travel. Sound is included to assist in this
process. From take-off on Hispanoamérica Airlines until dinner is served and paid for in Restaurante La Torre, the user learns much useful vocabulary that he would most likely encounter while traveling. The dialogues and vocabulary words were recorded by natives of Spanish-speaking countries.

In a part of this travel segment, the traveler steps off the airplane and encounters with the dilemma of hailing a taxi to his hotel. A short spoken dialogue is presented depicting a common conversation between a tourist and a taxi driver. Written scripts, both in English and Spanish, appear on the screen at the student’s request. The tourist then arrives at the hotel La Madrileña inquiring after a room for the period of one week. Scanned graphics are used to depict a hotel lobby on the screen. Vocabulary words referent to a hotel lobby are presented as buttons over such items as suitcases, keys and packages. When pressed, these buttons permit the user to both see and hear the Spanish equivalent of these articles.

Another segment of this program takes place inside of a Spanish restaurant. Buttons are strategically located over various articles in the scene. When depressed they give a close-up view of a table setting and menu. Using programming techniques previously described, both vocabulary and a restaurant dialogue are given to increase the student’s vocabulary bank.

Tests and quizzes which follow each section of newly presented material challenge the student’s comprehension of the dialogues and vocabulary words. These allow the user to identify which areas of study ought to be focused upon more intensely.

The CASA program is still in its developmental stages, but the feedback as to its effectiveness is encouraging. With additional segments, CASA could become a comprehensive introduction to Spanish. Supplementary lessons will teach information such as additional verb tenses and conjugation in both regular and irregular verbs. The HyperCard software, with virtually unlimited options, has the capability of supporting these additions to Computer Assisted Spanish Applications. As this and other language instruction programs continue to increase in both popularity and availability, it is our hope that CASA and like software will be used more extensively in foreign language educational curricula.
YIDDISH AND ASL:
A LOOK AT LINGUISTIC PREJUDICE

Danial Parvaz
Brigham Young University

eth·no·cen·trism (eth'no-sên'trîz'am) n. 1. Belief in the superiority of one's own ethnic group.
   - From The American Heritage Dictionary of the English Language

Ethnocentrism - the very word leaves a bad taste in our mouths. The remaining flavors remind us of Hitler's xenophobic campaign against the Jewish people, blacks having second-class citizenship in South Africa, and white supremacists burning crosses on the lawns of minority citizens. Whenever one ethnic group considers itself to be superior to (and therefore more deserving of the good things in life than) another group, the result always has the same flavor - that of bitter tears, wounded hearts, and broken bodies. In recent times, there has been a consciousness-movement aiming to cure the primitive kill-or-be-killed instinct, replacing it with what Covey calls the "abundance mentality"\(^1\) - the realization that the world is big enough for all of us.

Yet ethnocentrism wears other guises. One of those guises is the myth of linguistic superiority. This is the notion that somehow the language of one group is superior to the language of another. "A belief," says The Cambridge Encyclopedia of Language,

that some languages are intrinsically superior to others is widespread, but has no basis in linguistic fact ... The view of modern linguistics is that a language should not be valued based on the political or economic influence of the speakers.\(^2\)

Two languages that have suffered due to the superiority myth are Yiddish and American Sign Language (ASL). Both languages have suffered accusations that they are not "real" languages, both have

\(^1\) Covey, S. Seven Habits of Highly Successful People. Cassette Seminar.

been marked as being "detrimental to thinking", both have had efforts made against them to stamp out the undesirable languages. Only in recent times have efforts been made to understand the nature of the languages.

At the outset it needs to be mentioned that both ASL and Yiddish are full, rich languages in their own rights. They possess morphology, phonology and syntax -- the essential "building blocks" of language. Both languages have rich, dynamic vocabularies which serve the needs of their language communities, and as we shall see, the languages in question have received the stamp of approval from the academic community.

It is a human tendency to deny the existence of things which, if admitted, would compel a person to change his or her belief system. Next to entropy, inertia is the ruling law in the universe. If speakers of language A can show that language B is in fact not a language, then the question of linguistic superiority is moot - and that is exactly what was tried with Yiddish and ASL. In the early 1930s the *Encyclopedia Britannica* stated that

[Yiddish is] essentially a folk tongue. It eludes all strict grammatical analysis, though efforts are being made to bring about some system to its written form.3

In the nineteenth century, many epithets were used in reference to Yiddish, among them being: *hittserne taytsh* (wooden German), *kugl loshn* (pudding language), *shulhoyf-loshn* (synagogue-courtyard language), and *zhargon* 4. In the minds of the "purists", Yiddish was just corrupted German- the very name for the language is an Americanization of *jüdisch-deutsch* (Jewish-German). Such an offshoot from the pure German language must never be admitted as a real language, the scholars said. It must be remembered that J.G. Becanus (1518-72) stated that German was the language that Adam spoke in Eden, and that the Old Testament was translated from German to Hebrew some time after the Tower of Babel incident.5 The Jews were tinkering with the sacred.

---

3 Geipel, J. *Mame Loshn: The Making of Yiddish.* West Nyack. p.11
5 *Cambridge Encyclopaedia of Language.* p. 7
ASL has also had its detractors. To many people, both the uninitiated and the professionally ignorant, ASL is simply "English on the hands", a mere argot made up of manual signs and poor English grammar. In his autobiography, *Deafness*, the deaf English poet David Wright referred to the sign language used in his schoolboy days as a "dumb-crambo" - and so are all sign languages, according to the detractors of ASL. Poor analyses of ASL lead to the erroneous conclusions that

... ASL consisted of unordered, mimetic gestures and was incomplete, inferior, situation-bound and concrete. Such terms represented not only a lack of understanding of manual language, but an oral language bias, from which it is exceedingly difficult to escape.¹

The iconic nature of many ASL signs would lead one to the mistaken conclusion that ASL is simply a sophisticated pantomime system.² In the older literature, sign language was seldom referred to as *la langue gestuel*, but rather as *la mimique* or *les gestes*³ for *langue* implied "tongue". Without the tongue, there could be no language.

Language has been considered to be the wellspring of thought. Since language permeates our thoughts, our dreams, our reasoning processes, it was concluded that anyone who did not speak a "true" language must not have any intelligence. Parents who taught their children a non-language were doing their children a gross disservice - they were raising idiots. In 1786 David Friedländer wrote:

[Yiddish is] so wild and lawless that it cripples the minds of the children to make them incapable even of thinking straight. How could they expect their neighbors to respect them, when they didn't even speak like human beings?⁴

With ASL, the issue was mostly involved with the question of speech. In the area of deaf communication the two prevailing schools of thought are: 1) the oralists, who argue that speech and lip-reading are to be the *sole* means of communication (even among

²Camhridge Encyclopaedia of Language. p. 7
the deaf themselves!\(^1\), and 2) The manualists, who advocate the use of a signed system, whether or not speech was ever learned.

The chief argument brought against the use of sign language was what Laurent Clerc called the "God's breath flim-flam"\(^2\). The argument was that since God formed man in his own image, and since God could speak ("... and God said, Let there be light..."), in order for man to follow the creator's design, he had to speak. Amman, the Swiss founder of the "German method" - which was really the same old oralist method in a new package - summed up the proposition as:

> ...creatures formed in God's Image ought out of necessity to be able to speak, and in this respect resemble their Creator.\(^3\)

It is difficult to grasp the effect these kinds of statements had on the Jews and the Deaf. They made their languages abominations - dirty secrets to be tucked away in dark corners. Instead of systems for expression and communication, they became debasing elements, keeping man from his divine potential. Such corruptions could never be allowed to remain unchecked.

History is replete with people who have tried to rid humanity of "corruption" - self-appointed prophets determined to work towards the breeding of the perfect race. On the scale of a Hitler, the evil is glaringly clear. The Jews are systematically ostracized from society as a prelude to genocide. There are subtler means of slavery, however. During the Arab occupation of Persia, the Persian language was suppressed. Only the valiant efforts of a few poets and writers prevented the Arabic idiom from supplanting Persian. This is known to every Iranian schoolchild: if you proscribe the language of a people, that people will scatter and die. Cut off a generation from using their parental language, and you have killed that culture as surely as you would have by sending the members of that culture to Dachau. In the case of Yiddish, it was simply fuel that was added to the fire of anti-semitism: the Yiddish language hampers thought, therefore Jewish people are deficient, therefore they are harmful to

\(^{1}\) Schein, J. *Speaking the Language of Sign*. New York, Doubleday. 1984. p.60
\(^{3}\) Ibid.
racial purity. Perhaps the sentiments against ASL and Yiddish are the results of xenophobia.

Alexander Graham Bell was a eugeneticist. He was born of "older" American stock (i.e. English or German) - a group who demanded restrictive immigration to the States, language uniformity, and eugenic reduction of the "unfit" in American schools. The deaf conveniently fell into the category of "unfit" for the reasons described above, in addition to the fact that they did not have the use of their ears. He viewed Sign Language as a "useless and pernicious" device. From the money Bell made by inventing the telephone (an accident - he was trying to invent a hearing aid for his deaf wife) he aided the cause of the oralists. Backed by nearly unlimited wealth, these "benefactors" obtained the power they needed to all but expunge ASL from the school system.

In the end, both attempts at suppression failed. The languages and the peoples survived, though the languages sustained grievous injuries. Although most enlightened institutions accept Yiddish and ASL as languages in their own right, there are many others who, for bureaucratic or other reasons, deny this truth. The most outdated sources are quoted to sustain their point of view. Any research produced contrary to their opinions is dismissed as poor scholarship and partisan.

For there has been research, new and exciting. In 1959, Max Weinreich's Gesikhhte fun der yidisher shprakh (History of the Yiddish Language) is a masterful work describing the evolution and structure of Yiddish. Solomon Birnbaum's Yiddish (1979) is also a scholarly work about the Yiddish language. ASL research began in 1965 with William Stokoe's Dictionary of American Sign Language based on Linguistic Principles - the first dictionary based on a phonological analysis of ASL. At the Salk Institute, Ursula Bellugi's research showed ASL to be full, rich language in its own right with poetry, humor, and other qualities that spoken languages possess. The validity of Yiddish and ASL has been demonstrated beyond all

---

1 Speaking the Language of Sign p. 63.
3 A friend of the author, on conversation with a Dean in the College of General Education at BYU, heard the Dean remark that "deaf people do not have a culture worth serious intellectual consideration."
doubt to linguists. The opinion of a small group of scholars, however, is unlikely to change centuries of prejudice in a hurry.

Is prejudice being overcome? The open-mindedness professed by most Northern Americans is easily challenged when faced with a foreigner whose command of English is less than fluent. We quickly become impatient with another's fledgling attempts to speak a language we have spent a lifetime practicing. The overt racism once evident in America is now replaced with a more covert linguistic discrimination. Out in the world there are people capable in every skill but spoken English, and they are constantly denied decent employment because of the "poor impression" that is made at the job interview. A case in point: two fully qualified deaf technicians work for a local Utah software company. When customers see that the technicians are using sign language, the customers no longer feel safe leaving their computer equipment with the technicians. There exists an impression that if the English used by a speaker is less than perfect, the speaker must in some way be deficient. We live in a country where mediocrity is tolerated, as long as we understand the mediocrity in fluent English.

Yiddish and ASL are not the only languages that have been in this plight-English was once considered a language fit only for the uneducated\(^1\). There are scores of instances where opinions are passed on the expressive media of whole cultures. If we wish to consider ourselves as a free nation, we must accept a more tolerant view of the languages found in this country today. Only then can we deserve the title of "a kinder, gentler nation."

---

\(^1\) *Speaking the Language of Sign* p.8. Thomas More was a valiant defender of the English Language: "That our language is called barbarous is but a fantasy..."
TEACHING A MAN TO FISH

Wendy B. Taylor
Brigham Young University

You are probably familiar with the Chinese proverb "Give a man a fish and you feed him for a day; teach a man to fish and you feed him for a lifetime." I would like to talk about developing student self-reliance through teaching survival English.

First, I would like to introduce myself and give you some of my background so that you can see my roots. In 1977 I graduated from BYU in home economics education. After teaching home economics for three years, I was called on an LDS mission to Hong Kong. One of my assignments there was to teach English in the Vietnamese refugee camps. Near the end of my mission I was transferred to the Thailand mission to teach English and cultural orientation to Vietnamese, Khmer, Laotian and Hmong refugees preparing to enter western societies. Since I have been home from my mission I have taught ESL in elementary school, high school and adult education programs. You can see that I come to ESL with a very home- and community-oriented frame of reference.

Many teachers that are trained to teach English as a Second Language are trained to prepare students to enter a university. There is a place for such programs. The Linguistics Department at BYU is geared towards training teachers in this way. There is also a need to train teachers to teach survival English.

My definition of survival English is "the English necessary to live within an English-speaking community." Before presenting any material a teacher should always ask himself, "Will this material help my students to live more comfortably in our community?"

While planning lessons for survival English, there are three basic principles to consider. I will discuss each of these principles in more depth.

1. Use the "real thing."
2. Give the students confidence.
3. Teach the students how to use basic resources.

Use the "Real Thing". I like to use realia within the classroom. I often substitute at Provo High School's adult education program. I always like to have the teachers tell me in advance what the topic of the lesson will be. Then I set to work gathering food packages, a suitcase full of
clothing items, cleaning supplies, medicine bottles or whatever is appropriate for the topic of the evening.

I feel that handling real items is a critical element of teaching survival English. For example, often students do not have practice reading labels from actual packages. When I first returned home from my mission, another sister requested that I visit a refugee family in my hometown. After introductions and small talk, the mother came to me carrying an aerosol can. The sons translated as she asked if the can contained bathroom cleaner. Bathroom cleaner? No, it was leather cleaner. How had she picked up the wrong item in the store? She had gone shopping in our large, well-stocked supermarkets. Not being able to read English she picked up the first aerosol can that she saw, not realizing that many different products come in aerosol cans. This is just one of many different examples that I could cite of problem situations resulting from lack of practice in survival English.

One of my responsibilities in the adult education program at Provo High School was to develop learning kits for practice with survival English. One of these kits contained fifteen to twenty prescription medicine bottles with all of the cautions common to such medicines. The students enjoy practicing English with the real bottles and I am often amazed at misconceptions had by even advanced students.

Another way to use the "real thing" in the classroom is to take the students out into the community. We sometimes do not realize that places we accept as part of our everyday lives are new and strange to newcomers in our culture. I have taken students on fieldtrips to the dentist office and the fire station here in Provo. Many other places would be good sites for fieldtrips, such as employment offices and the emergency room at the hospital.

Sometimes it is not feasible to take the students to the "real thing" in the community so sometimes we need to bring the community to the students. One way this can be done is by having guest lecturers come to class. When the child restraint law passed the state legislature, I invited an employee from the county health department to visit our classroom. He showed the different types of car seats and how to use them. He also showed a film stressing the importance of using child restraints while travelling. This experience promoted much discussion among the students. Hopefully, it influenced them to be among the few residents of the state to follow the child restraint laws.

Another way to bring the community into the classroom is to bring tapes of actual radio spots or telephone conversations. Of course there are commercial tapes available for classroom listening practice, but I have found
that bringing self-made tapes heightens student interest because they have a need to relate to the community around them.

Advertisements or community information can be taped directly from the radio.

When Joan Morley spoke at I-TESOL several years ago she suggested taping actual telephone conversations. At that time I purchased a wire-tap device which I have found very useful in obtaining clear recordings of two-sided telephone conversations. Of course, when making a recording it is necessary to inform the other speaker of your intent to tape the conversation. In order to not cause the speaker to alter his natural speech I have finished the conversation first, then asked for permission to use the completed tape. One of the best resources of listening practice material through the telephone is pre-recorded messages such as those used by many businesses. These businesses include theaters, points of tourist interest, bus companies, etc. The extension service, which is sponsored by the county, Utah State University, and the U. S. Department of Agriculture, often has tapes of various homemaking and gardening messages which can be accessed over the telephone.

*Give Students Confidence.* When I am teaching survival English I am often surprised at the complexity of the dialogs presented in the beginning level textbooks. While a textbook is a good resource for ideas, I almost invariably simplify important utterances to create what I call **BASIC frame sentences**. I capitalize and underline "basic" here because I want to stress how simple these sentences can and need to be.

For example, when planning to teach the topic of housing a teacher needs to think to himself, "What utterances related to this topic do my students absolutely need to express?" A student will probably need to speak with a landlord and various repairmen. He will likely need to tell about broken appliances or fixtures in his home. Thus, a basic frame sentence for housing could be:

The _____ is _____.

The student could learn the vocabulary to fill in the blanks. In this case, he could create the following sentences, among others:

The heater is broken.
The stove is broken.
The refrigerator is broken.
The window is broken.
The pipe is leaking.
The toilet is leaking.
The roof is leaking.
The faucet is leaking.

So the learning of one basic frame sentence with accompanying vocabulary allows the student express many needs.

I use the following basic frame sentence for teaching the lesson on emergencies.

```
Please send _____ I need help!
```

The vocabulary used to fill in the blank would include "a fire engine," "an ambulance," "the police." If use of the articles seems too difficult for the students I teach them to utter the expression without them. In fact, all that the student really needs to know how to do is to dial the telephone number 911, and say, "Ambulance!" or "Fire!" It would be advantageous if his basic frame sentences also included passive vocabulary such as "name," "address," and "phone number," so that he could respond to the dispatcher's questions. However, the telephone call could be traced through electronic equipment if the caller could be trained to stay on the phone until help arrives.

How does the use of basic frame sentences help the students to gain confidence? By learning a very simple utterance, rather than trying to memorize a long dialog, I feel that a student can remember how to express his need when he is faced with an urgent situation.

Often a young member of the family acts as a translator for his parents or other older relatives. However, sometimes the older, often less-fluent person is left to face an emergency alone. Even if a young person is there to translate at the doctor's office, for example, the older person feels more comfortable if he has a basic understanding of the English being exchanged, rather than having to depend solely on his child's translation.

When I first thought of the concept of creating basic frame sentences, I tried to think of sentences for each topic covered in survival English classes. For the topic of food and shopping I came up with the following sentence:

```
a I need _____ some
```

I envisioned that the students would use names of grocery items to fill in the blank, creating "I need some milk," "I need a loaf of bread," "I need some onions," etc. As I contemplated this basic frame sentence longer, however, I
realized that it was completely unnecessary. Students would generally not need to express their needs at a supermarket; they would simply obtain the desired items without using any language. (Of course if the supermarket is disorganized they might ask for location of items, i.e. "Where is the _____?") The moral to this story is: if you are going to create basic frame sentences, be sure that the student needs to use them.

Teach Students How to Use Basic Resources. Many basic resources are available that native speakers of English often take for granted. These include the telephone book, bus schedules, the library, and the newspaper. Often examples how to use these resources are available in textbooks that we use with the students. However, how much more meaningful would our lessons be if we used actual bus schedules from UTA or the Provo phone directory?

I teach students how to use the different parts of the phone book. They are often familiar with the white pages (only needing practice using alphabetical order), but have not used the blue pages which contain government listings and other valuable information. Students also need practice using the yellow pages. Suppose as a new member of a community that you need to find a family doctor. Where would you look in the yellow pages? Under "Doctors?" No, the listings are under "Physicians."

A few months ago a friend was complaining to me that he had a meeting in Provo at the same time that his wife had a doctor’s appointment in Orem. How could they do both, given that they only had one car? "Take the bus!" I suggested. "There is no bus from where I live," was his answer. I knew where he lived and I knew that several buses pass by there each hour. "She wouldn’t know how to make a transfer to get out to Orem," was his next excuse. I was surprised. This man came from a large city in Latin America where mass transportation is probably common. He was well-educated and had nearly perfect English. I am not acquainted with his wife and I may be making a false assumption that her English is as good as his. However, even if her English is at a lower level, she should have still been taught how to read a bus schedule in order to survive more comfortably within the community.

When teaching ESL in an elementary school in Salt Lake City, the other teachers and I made sure that the students had an opportunity to go to the city library, which was within walking distance from their homes, to obtain library cards and books. We felt that this was important because if the children checked out books the parents, often too busy to use the library themselves, would have books in the home. They would need to answer the children’s questions about the meaning of unfamiliar words. Hopefully, this would help
their English skills in an indirect way.

The newspaper is a wonderful resource for an English teacher. Of course, many reading activities can be drawn from its pages. I particularly like to teach use of the want ads during a unit on housing. In order for the students to successfully use want ads we need to teach them the abbreviations used, which I myself am often at a loss to decipher.

Other Resources. Many other resources are available in a community that can aid the teacher of survival English. Of course these will vary with the community, but the teacher should become aware of as many resources as possible.

City government is a good resource. We have taken our students on field trips to the fire station. Paramedics have also come to the classroom to give basic first aid demonstrations. Students that are in the senior citizen age group might be interested in participating in various services offered by government to senior citizens.

County governments offer the health departments and the extension service. Health department personnel came to our class to give the lecture on child restraints seats previously mentioned. The extension department was able to obtain seedlings of Oriental vegetables that had been developed for the Utah climate. Receiving the seedlings was an extremely popular activity with our Indochinese refugees.

Other government agencies are both resources and entities with which to interact. If our students do not work with the Immigration and Naturalization Service, they must work with Job Service. And most must deal with the IRS. Helping students to deal with common government forms is part of survival English. Agencies such as the IRS are usually willing to send out a worker to teach students how to fill out their forms.

The telephone company has a telephone that they loan for educational purposes. It has a real dial tone, buttons or a dial that really work and a busy signal. People that have lived in a foreign country know that one of the most frightening settings for using the target language is over the phone, since visual feedback is no longer present. Most ESL students need much practice with telephone conversation and need to be taught common telephone etiquette expected in our culture.

Utah Power and Light or other power companies offer consumer services that could be useful to the students. These include cooking demonstrations and demonstrations on improving energy conservation in the home.
Local dentists, doctors and the hospital are willing to come to class or to give tours of their offices. Students should be taught how to find the emergency entrance to the hospital and how to fill out the forms required there. They should also be taught other options for emergency care.

Utah Transit Authority is a good resource for students to learn about. Besides providing bus schedules, they also have a computerized telephone system that provides information about bus arrivals.

I have found that Deseret Industries and other thrift stores are excellent places to find copies of used magazines reasonably priced. Magazines are good sources of pictures for a picture file, especially magazines like National Geographic which could provide pictures about the homelands of the students, a good starting place for a discussion of cultural contrasts. A good friend of mine, a master teacher, drew upon such pictures as well as other pictures depicting different cultures in order to have the students "teach" her about their homelands. While she was learning culture, the students were highly motivated to speak English.

Adapting textbook lessons requires time and extra effort, but also rewards the teacher with higher student interest, motivation, and progress in adapting to the community.
FOR THE WANT OF A NAIL: SHOULD WE TEACH PRONUNCIATION TO TEACH GRAMMAR?

Alan Carter Covell
Brigham Young University

Introduction

As teachers of English as a Second Language, we are all familiar with the problems of teaching English grammar to non-native speakers. The question is, how can we teach grammar more effectively? The poor efficiency of existing grammar-dominated teaching methods has led to the development of audiolingualism, notional-functional approaches, and all the present communicative approaches to teaching English. However, the perfect teaching method still eludes us, and will no doubt continue to do so for some time (Blair, 1982).

Rationale

Most of the world's ESL or EFL students rely on a formal, grammatical approach (or "propositional" as described by Breen, 1984) to learning English. Organized grammar in a propositional syllabus is a nice, neat, easy-to-arrange teaching method, (from simpler to more complex structures) and it is easy to test and grade. Until someone achieves satori in teaching methodology, in a way that will fulfill all the above academic-bureaucratic requirements for syllabus design, grading, and testing, grammar teaching will be with us in ESL and EFL. Nevertheless, something is missing from the learning equation, for the results of our labors are usually not those we desire (Garret, 1986:134). Grammar teaching is often overdone, creating a distasteful attitude towards the language being learned (Blair, 1982:214). The orderly learning process prescribed by propositional grammar teaching, that language will be gradually assimilated, is not evident; not many learners can successfully transfer mechanical grammatical abilities to real world tasks (Taylor, 1987: 55).

The strange interlanguage fossilization patterns that appear in adult learners' speech (Selinker, 1972: 215) seem to indicate that the Brown studies of first language morphological acquisition (1973), which suggest invariant, increasing levels of complexity, have much less bearing on second language acquired competence for adults (Larsen-Freeman, 1975: 417). Second language learners often have gaps in their grammatical competence, using a morpheme higher on the order of acquisition with greater facility than a morpheme lower on the order of acquisition. Much the
same result is apparent with grammar teaching; the learner can at times use more complex structures more easily than simple ones. The lack of success with grammar teaching at teaching integrative skills is not due to lack of experimentation with teaching methods, grammar teaching methods abound. Grammar has been taught:

...inductively and deductively, in the native language and the target language, with explanation and without, in paradigms and in dialogues.

(Garret, 1987:134)

Thus, it is obvious that hidden factors are negatively influencing the effectiveness of grammar teaching.

Relevant Literature

Until the publication of the Dulay-Burt studies (1974) of child second language acquisition, native language interference was accepted as the basic cause of learning problems in foreign languages. However, the L1=L2 hypothesis of Dulay and Burt seemed to indicate that morphemes are learned in an invariant order by all child learners of English as a Second Language. Their 1974 article stated that "the sequences of acquisition of 11 functors obtained for Spanish and Chinese children are virtually the same." While the overall statistical picture may indicate this, Figure One from the Dulay-Burt article is article indicates several areas of noticeable disparity in levels of competence between the two native language backgrounds. These represent areas of linguistic commonality or dissimilarity, which are classified in this paper as:

A) [+] or [-] morphological in relation to English. The native language of the subject does or does not have the same type of morphological structure as English in this area (e.g. plurals).

B) [+] or [-] phonological in relation to English. The native language of the subject does or does not have the same phonological structure in that particular position (e.g. word-final) in relation to English.

The most notable disparities in achievement for the two groups in Figure One are in areas which are both [-] morphological and [-] phonological for the Chinese learners in relation to English. This can be seen in the differences in the group means for the plurals, long plurals, and third person singulars. (Chinese has no final alveolar fricatives, nor does it use the same morphological strategy for plurals or possessives; also third person

185
singular verb inflection is absent). When taken as a group or individually, these final fricatives as morphological markers demonstrate greater disparity of competence than any other group or single category, excepting the slightly larger "past regular" category, which is also [-] morphological and [-] phonological in Chinese.

Figure One  This figure from the 1974 Dulay-Burt article, with mean scores highlighted

The Ss in Dulay and Burt's group are not carefully ranked by age, length of exposure to a naturalistic English environment, or amount of formal instruction, any one of which might have affected the outcome of the analysis. However, the differences in acquisition of the morphemes, while significantly consistent within language groups in order of acquisition, is very different in levels of competence. Does this mean that children within the pre-pubescent Critical Period of Lenneberg (1967) are having troubles with transfer? Could this be evidence of phonological filtering?

Bloomfield (1933) addressed phonological filtering as a significant barrier to language learning and stated that people tend to ignore sounds that are non-phonemic in their native language (e.g. an English speaker ignores the phonemic nature of Chinese tones). Flege (1981) addressed the issue of foreign accents in children and adults, and categorically stated that there is no conclusive evidence adults are permanently constrained in the acquisition of
phonology by neurobiological factors. While production is important, the other side of the coin in learning pronunciation is perception, and without perception, there can be no consistent production (Leather, 1983). This has tremendous implications for the acquisition of second languages, for a non-salient or non-phonemic phone in the second language may be completely ignored by the learner, especially if there is no comparable morphological modifier demanding a sound-shift in the native language. This is the sort of phenomenon that I believe appears in the disparities of competence shown by Dulay-Burt's studies. What caused the Chinese children, all at a prime age for language acquisition (well within the Critical Period) to fall so far behind their Hispanic counterparts in some ways?

Raymond Baird's study (1973) indicates that morphophonemic competence in children learning English as a first language doesn't transfer across morphological boundaries. In other words, just because a child can handle the /s/, /z/, and /zl/ morphophonemic rules for the plural, does not grant the child immediate competence with the identical rules for possessives. While the Dulay-Burt study did not address pronunciation, this might explain why the Chinese children showed less difference in competence in [+ ] morphological areas, even with [- ] phonological features, such as contractible copulas.

My question at this point is why do some students seem to have so much difficulty acquiring sibilant-final English morphemes? Surprisingly, the Chinese Ss' in Dulay and Burt's study showed contractible copula, past irregular, and contractible auxiliary competence levels that reached closer to those of the Hispanic Ss (see Figure One). Of the Asian languages, Korean, Japanese, Chinese, Thai, Vietnamese, and Laotian all lack final sibilants. If the issue of phonological filtering is addressed as significant in reinforcing transfer in the acquisition of grammatical morphemes, a two level hypothesis is possible:

1) If the learner's native language is [+ ] morphological and [+ ] phonological in relation to the second language morpheme being learned, it will be easier to acquire the structure than for a learner whose language is [+ ] morphological and [- ] phonological (or vice-versa).

2) If the learner's native language is [+ ] morphological and [- ] phonological in relation to the second language morpheme being learned, it will be easier for that learner to acquire the structure than for a learner whose language is [- ] morphological and [- ] phonological.
If the above hypotheses are accepted, this would appear to indicate that if the learners don't have a phonological "hook" upon which to hang a morphological "tag," they may never acquire the structure unless it is repeatedly reinforced so that the phonological filtering is overcome.

Methodology

The Ss in this study numbered twenty-two, all from the fourth level of instruction at the English Language Center at Brigham Young University. All were within the eighteen to twenty-five age category. Ten were native speakers of Japanese; twelve were native speakers of Spanish. The testing instrument (Appendix A) was a listening comprehension test, with one contextual /s/ differentiation, and six plural contextual differentiations. These consisted of three final /s/, two final /z/, and two final /z/. The Ss took the test by listening to the sentences read aloud, then chose a correct answer from three written choices. The native-speaker instructors from each class took the test along with the Ss as a control measure; if one question had been missed by a native speaker, it would have to have been deleted from consideration. This did not occur.

Besides the answers, basic scholastic data about the Ss were elicited, such as years of English in their native country, English in the U.S.A., and native language background. These were included as independent variables in the analysis. The answers were scored according to perceptual salience, with /s/ being least salient (devoiced) counting three points, /z/ being next most salient (voiced) counting two points, and /z/ being most salient (voiced and vowel reinforced) counting one point. The scoring hierarchy was determined by level of audibility, with the idea that voicing was louder than devoicing, and an added vowel added duration (and thus greater noticeability) to the sound. This accorded a possible total score of fifteen.

Analysis

Three one-way analyses of variance were performed to determine whether or not significant differences separated the groups of Ss when grouped according to the independent variables. The Ss were analyzed according to native language background, English in the U.S.A., and English as a foreign language. Since the groups being tested are Hispanic ([+] morphological and [+] phonological in relation to English plurals), and Japanese ([+] phonological and [-] morphological in relation to English plurals), the results were held significant if p<.01. See Table One for the results of this analysis.
### Table One
Analyses of Variance by ESL, EFL, and Total Score

#### Amount of English in the U.S.A. (years)

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Sd</th>
<th>DF</th>
<th>Sum/sq</th>
<th>Mean/sq</th>
<th>F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPN</td>
<td>10</td>
<td>.645</td>
<td>.312</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPN</td>
<td>12</td>
<td>.525</td>
<td>.259</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>1</td>
<td>0.079</td>
<td>0.079</td>
<td>0.973</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Groups</td>
<td>20</td>
<td>1.615</td>
<td>0.081</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>1.693</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pr&gt;f</td>
<td>0.663</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Amount of English in Native Country (years)

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Sd</th>
<th>DF</th>
<th>Sum/sq</th>
<th>Mean/sq</th>
<th>F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPN</td>
<td>10</td>
<td>6.70</td>
<td>1.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPN</td>
<td>12</td>
<td>2.33</td>
<td>1.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>1</td>
<td>104.01</td>
<td>104.01</td>
<td>38.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Groups</td>
<td>20</td>
<td>54.27</td>
<td>2.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>158.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pr&gt;f</td>
<td>*0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Total Score on Perception Test

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Sd</th>
<th>DF</th>
<th>Sum/sq</th>
<th>Mean/sq</th>
<th>F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPN</td>
<td>10</td>
<td>1.90</td>
<td>1.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPN</td>
<td>12</td>
<td>6.58</td>
<td>3.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>1</td>
<td>119.64</td>
<td>119.64</td>
<td>16.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Groups</td>
<td>20</td>
<td>143.82</td>
<td>7.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>263.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pr&gt;f</td>
<td>*0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**JPN** = Japanese native language group  
**SPN** = Spanish native language group

### Findings and Discussion

As seen in the results of the analyses of variance and in Figure Two (below), there is no significant difference in the amount of English in the United States between the Japanese and Hispanic groups, while there is a significant weighting towards the Japanese Ss (p<.001) in the amount of English as a Foreign Language. It would seem reasonable that extra instruction should provide advantages for the Japanese Ss in being tested in a simple grammar task, in this case identification of the plural morpheme. This is manifestly not the case. It is apparent that the "grammar
"teaching" did not provide any benefit at all for the Japanese native language group in an integrative listening-grammar skills task.

Regardless of the greater amount of EFL for the Japanese Ss, the Hispanic native-language group did better on the final-sibilant perception test at the p < .001 significance level than did the Japanese Native-language group. The Hispanic group had a significant advantage; this seems directly attributable to phonological filtering by the Japanese group. All the Ss are from the same level of competence, and had comparable amounts of ESL instruction.

Figure Two

Japanese & Hispanic ESL Students' Perception of Final Sibilants

Mean Raw Score

![Graph showing the perception of final sibilants by Japanese and Hispanic ESL students. Series A represents Japanese Ss, Series B represents Hispanic Ss. The graph indicates that Series B (Hispanic Ss) performed better with a mean raw score of 3.28 compared to Series A (Japanese Ss) with a mean raw score of 1.96. The significance level is p < .001 for ESL and p < .001 for EFL.]
Conclusions and Suggestions for Further Research

While the testing instrument in Appendix A needs further refinement, and other sibilant-final morphemes examples to see if this result holds true for other morphological concepts as well, this pilot study indicates that there may be a significant reinforcing phonological-morphological acquisition blocking phenomenon at work here. If this is the case, any learner from a native language that is [-] morphological and [-] phonological in relation to a given grammatical morpheme in English will have a much more difficult time with that particular structure than a student from a native language that is [+ ] morphological and [+ ] phonological in relation to that structure.

Further tests of this concept with a third group ( [+ ] morphological and [- ] phonological or [- ] morphological and [+ ] phonological group in relation to a given English morpheme) to balance the extreme [+]/[+] to [-]/[-] linguistic separation between the Japanese and Hispanic Ss should also be done. This would test all possible levels of the hypotheses simultaneously.

This type test need not be limited to English. Tone-phonemic language learners from tone non-phonemic language backgrounds (Americans learning Chinese) should be contrasted with tone-phonemic native language background learners learning a second tone-phonemic language (such as a Thai learning Chinese). If the two groups of Ss are at the same levels of competence more phonological filtering may appear for the American group.

The lower disparity of results between the Chinese and Hispanic students in Figure One (e.g. contractible copulas) even when the phonological tag is missing for the Chinese Ss may prove to be a vital clue. If the morpheme is easier to acquire when it is a shared concept, this may indicate shared morphology can overcome phonological filtering.

What factors will affect changes in this problem in relation to the teaching of grammar by using pronunciation more successfully one can only hypothesize; procedures might include the repeated stressing of perception and production of novel morphemes for students who don't have it in their morphological or phonological inventory.

Which is more important, the phonological sound or the morphological concept, is a veritable "chicken or egg" question, and as yet unanswerable; but there is one certainty. Until those learners who filter sounds hear them, they may never develop a morphological relationship to fit the phonetic occurrence, no matter how much grammar they study. Without the teaching of pronunciation to foster
perception, the language acquisition device may never be able to sort out the rules for spontaneous creative construction at a native-speaker level. For the teaching of grammar to be truly effective in modern language courses, it might be necessary to teach pronunciation and listening comprehension as an integral part of it.

REFERENCES


APPENDIX A

Test of final sibilant consonants

1. "His son was just six yesterday."
   A. His son was not ill.
   B. His son had a birthday
   * C. His son was ill only yesterday. 3 points /s/

2. "Where shall I put the books?"
   A. Put the book here.
   B. Is it very heavy?
   * C. Set them over there. 3 points /s/

3. "Get the boxes from the car."
   A. The box isn't there now.
   B. Which box do you want?
   * C. How many are in there? 1 point /z/

4. "Are you going to buy her the roses?"
   * A. Only if I can find them. 1 point /z/
   B. No, it is too much.
   C. Yes, if she wants one.

5. The boys want a to go to a movie.
   A. He can go tonight.
   B. Will he do his homework first?
   * C. What time will they come back? 2 points /z/

6. "Can I please get the gloves George has?"
   A. No, it's too big for you, isn't it?
   B. If George says so, you can get it.
   * C. Who will help you put them on? 2 points /z/

7. Our bikes should be put away at night.
   A. Don't forget to put it away tonight.
   * B. Someone might take the bikes from outside. 3 points /s/
   C. Will you help put the bike away?
PROGRESS REPORT
ON THE BOOK OF MORMON
CRITICAL TEXT PROJECT

Royal Skousen
English Department
Brigham Young University

(1) CURRENT STATUS OF THE 5-YEAR PROJECT (as of 7 July 1989)

Original Manuscript (O):

(only about 25% remains: 1 Nephi 2 \rightarrow 2 Nephi 1 / Alma 22 \rightarrow Helaman 3, with gaps; plus other fragments)

permission of First Presidency to make transcription of the original manuscript: photos on loan from historical department

fragment at University of Utah: photo purchased

work done:

preliminary first transcription by Royal Skousen
first transcription checked up to Alma 50
second transcription by Lyle Fletcher and Marcello Hunter
line-by-line comparison of two transcriptions through Alma 32

work to be done:

finish checking first transcription
finish line-by-line comparison of transcriptions
check transcriptions against actual manuscript
do remaining fragments
compare O against P and printed editions

Printer’s Manuscript (P):

(all extant except for 4 lines from 1 Nephi 1)

enlarged photographic reproduction on loan from RLDS Church Historian (Richard Howard)

work done:

first transcription by Royal Skousen through the small plates
second transcription by Lawrence Skousen
work to be done:

finish first transcription
make line-by-line comparison of two transcriptions
compare against O and printed editions
distinguish between the three correctors of P

Scanning of Important Editions of the Book of Mormon
(+ finished, - partially finished)

table:

<table>
<thead>
<tr>
<th>edition</th>
<th>copy</th>
<th>xerox</th>
<th>scan</th>
<th>format</th>
<th>proof</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1830</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>1837</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>1840</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>1841</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>1849</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1852</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1874</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1905</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1911</td>
<td>+</td>
<td></td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1920</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>1981</td>
<td>+</td>
<td></td>
<td>+</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>RLDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1858W</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1874R</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1892R</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1908R</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1953R</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1966R</td>
<td>+</td>
<td></td>
<td>+</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Publication Plans:

transcripts of O and P

critical text:

original text as dictated by Joseph Smith
(to the extent that it can be determined)
dictated format: limited punctuation,
but modern spelling and lexical capitalization
apparatus: all significant textual variants
(including all meaning differences
and some spelling and punctuation differences)
LDS versification (left margin)
RLDS versification (right margin)
cross-references to explicitly quoted biblical passages

195
(2) SIGNIFICANT ERRORS BETWEEN THE ORIGINAL MANUSCRIPT (O) AND THE PRINTER’S MANUSCRIPT (P)

The printer’s manuscript is a "copy" of the original manuscript, but as in all copying, errors enter in. In the printer’s manuscript, there is, on the average, one or two textual errors per manuscript page. Most of these errors do not make a difference in meaning, but about one in five errors do. In the following list of errors (based on a preliminary comparison between the two manuscripts), I provide an analysis of 39 significant copying errors (ones that make a difference in meaning), plus one interesting error made by the compositor for the 1830 edition. Of these 40 errors, 23 of them have been newly found in our critical text project (although RLDS researchers have independently found many of these). These newly found errors are marked with an asterisk.

In addition, I categorize the errors according to whether the reading in O is the harder or the easier reading. Most of the harder readings are found in P and are accidental copying errors:

h harder reading in O 6
e easier reading in O 23=
equivalent readings 11

Moreover, there are many more examples of where the copying led to a contracted text rather than an expanded one:

+ expanded text in P 1
- contracted text in P 11=
equivalent length 28

In the following list, the harder reading analysis is given first, followed by the text length analysis. The abbreviations 1n and 2n stand for the books of First and Second Nephi, while aa stands for the book of Alma. The relevant item is underlined in the original manuscript reading and is separated from the printer’s manuscript reading by the symbol >. A caret ^ stands for an insertion; a null symbol Ø stands for a deletion.

=,= 1n3:10 when we had gone up to the land of Jerusalem ∴ come

=,+ 1n7:1* but that his sons should take daughters to wife that ^ might raise up seed unto the Lord ∴ they
h,= 1n7:17  O Lord according to my faith which is in me > thee
e,= 1n8:31* and he also saw other multitudes pressing their way towards that great and spacious building > feeling
=,= 1n13:4 and it came to pass that I saw among the nations of the gentiles the formation of a great church and the angel said unto me behold the formation of a church > foundation, foundation
e,= 1n13:12* and I looked and beheld a man among the gentiles which were separated from the seed of my brethren > was
=,= 1n13:24 and when it proceeded forth from the mouth of a Jew it contained the fulness of the gospel of the Lord > plainness/
e,= 1n13:24* of whom the twelve apostles bore record and they bore record according to the truth > bear, bear
=,= 1n13:26 thou seest the formation of that great and abominable church > foundation
e,= 1n15:12 the house of Israel was compared unto an olive tree by the spirit of the Lord which was in our father > fathers
=,- 1n17:3 he doth nourish them and strengthen them and provide ways and means whereby they can accomplish the thing which he hath commanded them wherefore he did provide ways and means for us > \emptyset, \emptyset
h,- 1n17:50* if he should command me that I should say unto this water be thou earth and it shall be earth > \emptyset, should
e,= 1n18:7 and now my father had begat two sons in the wilderness the elder was called Jacob and the younger Joseph > eldest
=,- 1n19:2 wherefore the record of my father and the genealogy of his forefathers and the more part of all our proceedings in the wilderness are engraven upon those first plates > \emptyset
e,= 1n22:8* wherefore it is likened unto the being nourished by the gentiles > nourished
e,- 2n1:1 our father Lehi also spake many things unto them and rehearsed unto them how great things the Lord had done for them > \emptyset
we have obtained a land of promise a land which is choice above all other lands a land which the Lord God hath covenanted with me should be a land for the inheritance of my seed yea the Lord hath consecrated this land unto me > covenanted

their brethren the Lamanites made preparations for war and came up to the land of Nephi for the purpose of dethroning the king and to place another in his stead and also of destroying the people of Anti Nephi Lehi out of the land > destroying

in the commencement of the seventeenth year > Ø

for it was strictly contrary to the commandments of God > commands

and I always knew that there was a God > also

thanking their God that they were chosen of him and that he had not led them away > did, lead

how long wilt thou suffer that such wickedness and infidelity shall be among this people > iniquity

after much labor among them they began to have success among the poorer class of the people > poor, Ø

but Zenoch also spake of these things > Zenock

if ye could be healed by merely casting about your eyes that ye might behold would ye not behold quickly > be healed

yea I say unto you my son that there can be nothing so exquisite and so bitter as was my pains yea and again I say unto you my son that on the other hand there can be nothing so exquisite and sweet as was my joy > could

even so I would that ye should continue to teach and I would that ye should be diligent > would

and he was only twenty and five years old when he was appointed chief commander over the armies of the Nephites > captain

there was now and then a man fell among the Nephites by their wounds and the loss of blood > swords
Moroni and his army met the army of the Lamanites. Therefore he had power to do according to his will with the armies of the Nephites to establish and to exercise authority over them.

Amalickiah caused that his servants should go forth to meet the king and they went forth and bowed themselves before the king.

In the latter end of the nineteenth year notwithstanding their peace, the people of Nephi did thank the Lord their God because of his miraculous power in delivering them.

And Pahoran retained the judgment seat which caused much rejoicing among the brethren of Pahoran and also among the people of liberty.

And now it came to pass in the commencement of the twenty and ninth year.

And thus were we favored of the Lord for had they come upon us in this our weakness they might have perhaps destroyed our little army but thus were we preserved.

And it came to pass that I thus did send an embassy to the great governor of our land.

An error made by the printer:

He sent a petition with the voice of the people unto the governor of the land desiring that he should heed it and give him Moroni power to compel those dissenters.

A few observations:

1. The error rate is fairly constant, with about one significant textual error every two or three manuscript pages; since we have only about 25% of the original manuscript, I would estimate there are at least 100 textually significant errors in the missing portions.
(2) The errors are relatively minor, thus explaining why these errors entered the text; conjecturing will probably not be very helpful in discovering the unknown errors.

(3) Where textual evidence does occur, we can usually determine which reading is the more difficult one — and surprisingly the errors that have crept in tend to introduce more difficult readings (contrary to the usual assumption in textual criticism); the harder readings in 0 are generally defensible, the harder readings in P are not.

(4) The text does not tend to grow; in fact, contractions are much more frequent than expansions.

(5) The original manuscript is clearly the better text.

(6) Oliver Cowdery and the other scribes tried to make a faithful copy; there is very little evidence that they consciously expanded or tried to explain the text; nearly all errors are natural transcriptional errors.

(7) Only conscious editing leads to an expansive and easier text (as is found in many printed editions of the Book of Mormon from 1837 on).

(8) These results have important implications for biblical textual criticism. There are striking similarities between the textual development of the Book of Mormon and the scriptures of the early Christian church: faithful members with some education (but not trained in manuscript production) make (imperfect) copies; errors enter in from the beginning and are not recoverable if the original is destroyed. In this regard, Kurt Aland’s claims (The Text of the New Testament 291) regarding the New Testament text seem incredibly naive:

"any reading ever occurring in the New Testament textual tradition, from the original reading onward, has been preserved in the tradition and needs only to be identified"