NOOJ MORPHOLOGICAL GRAMMARS FOR STENOTYPE WRITING
A MULTILINGUAL METHOD FOR REAL-TIME AUTOMATIC TEXT CORRECTION

MARIO MONTELEONE – RAFFAELE GUARASCI – ALESSANDRO MAISTO
UNIVERSITÀ DI SALERNO
mmonteleone@unisa.it rguarasci@unisa.it amaisto@unisa.it

May 18-20, 2017, MISC Laboratory, Faculty of Science, Ibn Tofail University, Kenitra; National School of Computer Science and Systems Analysis (ENSIAS), Rabat, Morocco.
What is Stenotyping? - 1

- Writing method system used to transcribe spoken texts, rapidly and in real time
- Mechanical or digital device used: Stenotype, stenotype machine, short-hand machine or steno writer, equipped with a special keyboard allowing to performing beats of one or more keys simultaneously. Used in courts (especially in the USA)
- It requires the application of specific coded writing systems to limit and accelerate the number of beats
- In the USA, a trained court reporter or closed captioner must write speeds of approximately 180, 200, and 225 words per minute (some stenographers can reach 300 words per minute)
The Stenograph is used with a specially modified version of the “Metodo Melani”, worldwide one of the most used system for stenotyping.

A stenotype keyboard has far fewer keys than a conventional alphanumerical key-board (23 keys, the letter “S” to the left of the keyboard and the asterisk both are single keys).
What is Stenotyping? - 3

- Very large number of potential combinations:
  - factorial combinations: $23! = 25,852,016,738,884,976,640,000$
  - combinations per maximum number of fingers: $23^{10} = 41,426,511,213,649$
  - combinations for maximum average number of morphemes in an Italian word: $41,426,511,213,649/3 = 13,808,837,071,216,33$ (periodical)

- To spell out whole syllables, words, and phrases with a single hand motion, multiple keys are pressed simultaneously (a procedure known as “chording” or “stroking”)
The «Metodo Melani» - 1

-created for Italian in 1980 by Professor Marcello Melani (Castelfranco di Sotto, 22-11-1928, Pisa, 12-4-2012, scholar and teacher of shorthand and computer analysis)

- Directly compatible with electronic processing
- In the ‘90s, its author adapted it to also Spanish and Portuguese

“Developments in computer technology had enabled shorthand machine to take advantage of electronic processing for a stenotype automatic transcription, and can therefore realize the mirage of real-time writing, which has now become a reality, not theoretical but practical.”

Mario Monteleone Mario Monteleone – Raffaele Guarasci – Alessandro Maisto (UNISA - Italy)
Main feature of Melani’s technical system is the **full writing (not abbreviated)** of a text

Real-time production of verbatim computer-assisted reports “in view of the application processor”, taking into account the compatibility with the stiffness and potential offered by computers, especially in term of real-time support (much more important for Criminal Court trials than for Civil proceedings)

The computer input is relatively simple: there is no need for sophisticated algorithms or particularly large dictionaries of abbreviations, the computer simply recognize some shortened codes, conceived from the beginning so as not to give rise to ambiguities, and managing a dictionary of acronyms
The “Metodo Melani” disassembles the keyboard into two basic parts: the part operated by the left hand, and the part operated by the right hand.

In order to write correctly, it is necessary to break the words into syllables.
The «Metodo Melani» - 4

Left part, used to write consonants that occur as first letters in words or syllables (left-hand fingers in succession: little finger, ring finger, middle finger, index finger).

Middle-lower part used to write vowels that occur in the initial part of syllables or in the middle part of words (fingers in succession: left-hand thumb finger for the “i” and “a”; right-hand thumb finger for “e” and “o”; the vowel “u”, not displayed, stroking contemporarily “e” and “o” with the right-hand thumb finger.

Central-right part used to write consonants which occur in the middle or fine parts of words (right-hand fingers in succession: index finger, middle finger, and ring finger.

Right part used to write vocals occurring at the end of words (typed exclusively at the end of a word, even if composed by only one syllable. Right-hand little finger of the right hand is the only finger used. The vowel “u”, stroking contemporarily “a” and “o”, always with the right-hand little finger.)
For the letters not present on the Stenograph keyboard, it is necessary to use key combinations

(left part) for the consonants occurring at the beginning of words or syllables: $PTV = B; TH = D; TV = F; PC = G; HR = L; CHR = M; H = N; PTVCHR = Q$ (six keys simultaneously pressed exactly on the splitting slot); $SPT = Z$

(central-right part) for the median or final word consonants: $CTP = B; TH = D; TP = F; PR = G; HR = L; SHR = M; H = N; CTPXHR = Q$ (six keys simultaneously pressed exactly on the splitting slot); $CT = V; SH = Z$
The keyboard has also an autonomous number bar:

- Left part: “S + bar” for “1”; “P + bar” for “2”; “T + bar” for “3”; “V + bar” for “4”; “I + bar” for “5”
- Right part: “C + bar” for “6”; “T + bar” for “7”; “P + bar” for “8”; “I + bar” for “9”

The asterisk is used both to write the asterisk itself, when stroked contemporarily with the middle vowel “a”, both to delete the last word written, when the stenotype machine is connected to a computer.

The “Metodo Melani” is customizable (it allows the encoding of the most frequently used words, and their association to the stroking of specific keys)
The typing procedure is completely mnemonic and forces stenotype writers to remember the keys to stroke on the base of the positions the letters to write have inside words, and not and the base of their morphological values.

As for word segmentation, this method is not based on the identification and coding of morphemes, but at best of syllables:

- The subdivision of words into syllables does not return fixed minimum units (no specific lists of syllables exist a priori in any given language, and syllable forms and contents may vary according to the words hyphenated).

- The subdivision of words into morphemes restores fixed, stable, and reusable minimum units, definable thanks to the specific word-formation rules pertaining to each language (from inflectional to polysynthetic).
The speed of required beats often generates a high amount of typos

Any eventual automatic correction of typos must be completed during post-editing. It is possible to verify the correctness of each word written only at the end of typing

The “Metodo Melani” cannot be considered as a standard in computational/combinatory procedure or routines because its base elements:
- Are not taxonomically coded as for their functions
- Are not iteratively reusable
- Have a high degree of redundancy
The creation of a stenotype writing method based on a non-casual combination of morphemes would rely on a defined list of elements to be combined (i.e., the morphemes of a language) together with a production syntax (i.e., the word-formation rules of a language).

The morphemic structures of languages are more stable/predictable than syllables, and more easily to memorize and store, and process.

The creation of a morphological-based stenotype method may limit the write choices of operators, thus reducing the number of possible typos.
NooJ linguistic resources and morphological grammars can be used to build and implement a system for real-time typos automatic correction during stenotype writing.

Such grammars can be built to account for both lexical constellations (i.e., groups of words sharing the same lexical morpheme) and morpheme combinations inside words (i.e., by means of segmentation).

This system will be built as a supplementary resource to NooJ (Python Module?), to be embedded in operating systems and/or word processors of computerized shorthand machines.
1. **New subdivision of keyboard** on the basis of Italian implicit phonological rules

2. **Detection and classification of all the morphemes** of the Italian language (by means of DELAF entry segmentation)

3. **Association of each tag or sequences of tags, to one or more keys** of the Stenograph machine, also using keystrokes (not to write syllables or portion of words, but morphemes and attested sequences of morphemes, ie complete words)

4. **Factorization** of the writing rules obtained

5. **Use of Italian DELAF** (over two million entries) to **check, validate and eventually debug** the combinations produced
This method is **adaptable to any language** that can formalize its phonological and morphological rules.

**Phonological Hints** (keystroking by **minimal pairs**): Alveolars Fricatives: s/z; Bilabial Occlusives: p/b; Velar Occlusives: c/g; Labiodental Fricatives: v/f; Alveolar Vibrant and Approximant: r/l; Alveolar Occlusives: t/d

- The “p” in *pet* = **PTV**
- The “b” in *bet* = **PPV** (doubling one sound of the pairs)

This change will not increase the number of letters to stroke, and will follow the criteria of speaker’s intuition and phonological logics.
Morphological Hints (NooJ dictionary):

#Lexical morphemes
ahcr, LEX+type=angl
vrahc, LEX+type=franc
cerhah, LEX+type=german
crec, LEX+type=grec
itar, LEX+type=ital

#Derivational morphemes
ish, SFX+type=ism
ist, SFX+type=ist
ic, SFX+type=ic
hht, SFX+type=mente
iss, SFX+type=izz
apr, SFX+type=abil
itaa, SFX+type=itè
are, SFX+type=are
at, SFX+type=at
ash, SFX+type=azion
ovp, SFX+type=ofob
ovh, SFX+type=ofon
es, SFX+type=es
iah, SFX+type=ian

#Grammatical morphemes
a, GRM+type=a
e, GRM+type=e
i, GRM+type=i
o, GRM+type=o
he, GRM+type=he
hi, GRM+type=hi

The non-random combination of these 27 morphemes “produces” 161 usable Italian words.
<table>
<thead>
<tr>
<th>1. anglicana</th>
<th>40. francesismo</th>
<th>80. grecizzabili</th>
<th>120. italo</th>
<th>160. germanofonia</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. anglicanamente</td>
<td>41. francesistica</td>
<td>81. grecizzabilità</td>
<td>121. italofoba</td>
<td></td>
</tr>
<tr>
<td>3. anglicane</td>
<td>42. francesisticamente</td>
<td>82. grecizzare</td>
<td>122. italofobe</td>
<td></td>
</tr>
<tr>
<td>4. anglicani</td>
<td>43. francesistiche</td>
<td>83. grecizzata</td>
<td>123. italofobi</td>
<td></td>
</tr>
<tr>
<td>5. anglicanismi</td>
<td>44. francesistici</td>
<td>84. grecizzate</td>
<td>124. italofobia</td>
<td></td>
</tr>
<tr>
<td>6. anglicanismo</td>
<td>45. francesistico</td>
<td>85. grecizzate</td>
<td>125. italofobo</td>
<td></td>
</tr>
<tr>
<td>7. anglicano</td>
<td>46. francesizzabile</td>
<td>86. grecizzato</td>
<td>126. italofona</td>
<td></td>
</tr>
<tr>
<td>8. anglicismi</td>
<td>47. francesizzabili</td>
<td>87. grecizzazione</td>
<td>127. italofone</td>
<td></td>
</tr>
<tr>
<td>9. anglicismo</td>
<td>48. francesizzabilità</td>
<td>88. grecizzazioni</td>
<td>128. italofoni</td>
<td></td>
</tr>
<tr>
<td>10. anglicistica</td>
<td>49. francesizzare</td>
<td>89. greco</td>
<td>129. italofonia</td>
<td></td>
</tr>
<tr>
<td>11. anglicisticamente</td>
<td>50. francesizzata</td>
<td>90. grecofoba</td>
<td>130. italofono</td>
<td></td>
</tr>
<tr>
<td>12. anglicistiche</td>
<td>51. francesizzate</td>
<td>91. grecofobe</td>
<td>131. germana</td>
<td></td>
</tr>
<tr>
<td>13. anglicistici</td>
<td>52. francesizzate</td>
<td>92. grecofobi</td>
<td>132. germane</td>
<td></td>
</tr>
<tr>
<td>14. anglicistico</td>
<td>53. francesizzato</td>
<td>93. grecofobia</td>
<td>133. germani</td>
<td></td>
</tr>
<tr>
<td>15. anglicizzabile</td>
<td>54. francesizzazione</td>
<td>94. grecofobo</td>
<td>134. germanismi</td>
<td></td>
</tr>
<tr>
<td>16. anglicizzabili</td>
<td>55. francesizzazioni</td>
<td>95. grecofonia</td>
<td>135. germanismo</td>
<td></td>
</tr>
<tr>
<td>17. anglicizzabilità</td>
<td>56. franche</td>
<td>96. grecofone</td>
<td>136. germanistica</td>
<td></td>
</tr>
<tr>
<td>18. anglicizzare</td>
<td>57. franchi</td>
<td>97. grecofoni</td>
<td>137. germanisticamente</td>
<td></td>
</tr>
<tr>
<td>19. anglicizzata</td>
<td>58. franco</td>
<td>98. grecofonia</td>
<td>138. germanistiche</td>
<td></td>
</tr>
<tr>
<td>20. anglicizate</td>
<td>59. francofoba</td>
<td>99. grecofona</td>
<td>139. germanistici</td>
<td></td>
</tr>
<tr>
<td>21. anglicizate</td>
<td>60. francofoe</td>
<td>100. italia</td>
<td>140. germanistico</td>
<td></td>
</tr>
<tr>
<td>22. anglicizzato</td>
<td>61. francofobi</td>
<td>101. itale</td>
<td>141. germanizzabile</td>
<td></td>
</tr>
<tr>
<td>23. anglicizzazione</td>
<td>62. francofobia</td>
<td>102.iali</td>
<td>142. germanizzabili</td>
<td></td>
</tr>
<tr>
<td>24. anglicizzazioni</td>
<td>63. francofo</td>
<td>103. italianismi</td>
<td>143. germanizzabilità</td>
<td></td>
</tr>
<tr>
<td>25. anglismi</td>
<td>64. francofoa</td>
<td>104. italiano</td>
<td>144. germanizzare</td>
<td></td>
</tr>
<tr>
<td>26. anglismo</td>
<td>65. francofone</td>
<td>105. italiana</td>
<td>145. germanizzata</td>
<td></td>
</tr>
<tr>
<td>27. anglo</td>
<td>66. francofini</td>
<td>106. italianisticamente</td>
<td>146. germanizate</td>
<td></td>
</tr>
<tr>
<td>28. anglofoba</td>
<td>67. francofonia</td>
<td>107. italianistiche</td>
<td>147. germanizate</td>
<td></td>
</tr>
<tr>
<td>29. anglofobe</td>
<td>68. francofono</td>
<td>108. italianisti</td>
<td>148. germanizzato</td>
<td></td>
</tr>
<tr>
<td>30. anglofobi</td>
<td>69. greca</td>
<td>109. italiano</td>
<td>149. germanizzazione</td>
<td></td>
</tr>
<tr>
<td>31. anglofobia</td>
<td>70. greche</td>
<td>110. italianoizzabile</td>
<td>150. germanizzazioni</td>
<td></td>
</tr>
<tr>
<td>32. anglofobo</td>
<td>71. greci</td>
<td>111. italianoizabilì</td>
<td>151. germano</td>
<td></td>
</tr>
<tr>
<td>33. anglofonia</td>
<td>72. grecismi</td>
<td>112. italianizzabilità</td>
<td>152. germanofoba</td>
<td></td>
</tr>
<tr>
<td>34. anglofone</td>
<td>73. grecismo</td>
<td>113. italianizzare</td>
<td>153. germanofobe</td>
<td></td>
</tr>
<tr>
<td>35. anglofoni</td>
<td>74. grecistica</td>
<td>114. italianoizzata</td>
<td>154. germanofobi</td>
<td></td>
</tr>
<tr>
<td>36. anglofonia</td>
<td>75. grecisticamente</td>
<td>115. italianoizzate</td>
<td>155. germanofobia</td>
<td></td>
</tr>
<tr>
<td>37. anglofono</td>
<td>76. grecistiche</td>
<td>116. italianoizzate</td>
<td>156. germanofobo</td>
<td></td>
</tr>
<tr>
<td>38. franca</td>
<td>77. grecistici</td>
<td>117. italianoizzato</td>
<td>157. germanofona</td>
<td></td>
</tr>
<tr>
<td>39. francesismi</td>
<td>78. grecistico</td>
<td>118. italianizzazione</td>
<td>158. germanofone</td>
<td></td>
</tr>
<tr>
<td></td>
<td>79. grecizzabile</td>
<td>119. italianizzazioni</td>
<td>159. germanofone</td>
<td></td>
</tr>
</tbody>
</table>
Specific NooJ grammars will describe word segmentations, and check the correctness of the words written.

Any different sequence of letters not complying with those in the graph will not come to the end of it, being therefore a typo.

NooJ unknown words will be considered as typos or new dictionary entries.

The following graph shows the segmentation of the noun “italianistica” (Italian studies) plus of the two adverbs “italianamente” (in the Italian way) and “italianisticamente” (in the Italianistic way).
Overcoming the limits of the «Metodo Melani» - 7
Example of stenograph annotations in a text

Mario Monteleone – Raffaele Guarasci – Alessandro Maisto (UNISA - Italy)
Overcoming the limits of the «Metodo Melani» - 9

NooJ dictionary of the annotated text

Mario Monteleone – Raffaele Guarasci – Alessandro Maisto (UNISA - Italy)
Morphemes will be typed by means of their corresponding strokes, and following their specific sequences: in *Germanizzabilità* (Germanizability), “german-” will be written stroking contemporarily the keys CERHAH:

\[
\begin{align*}
\text{ISS} & : s p t v * c t p i \text{A} \\
\text{APR} & : s p t v * c t \text{P i a} \\
\text{ITA} & : s p t v * c T \text{p i A} \\
\end{align*}
\]

The additional successive morphemes will be written in sequence:

\[
\begin{align*}
\text{ISS} & : s c h r * s \text{H R e o} \\
\text{APR} & : s c h r * s \text{H R e o} \\
\text{ITA} & : s c h r * s \text{H R e o} \\
\end{align*}
\]
Syntactic grammars/transducers (as the one of the next figure) will use the results of the annotation procedure to rewrite Stenograph key sequences into Italian words.

The numbered variables in the graph are to be considered as slots (types), which may be full or empty, depending on the morphological complexity of the words accounted for. In this case, the words rewritten are:

- **italianista**, **italianistica**, **italianisticamente**, **italianamente**, **germanamente**, **germanista**, **germanistica**, **germansiticamente**
Conclusions and Further Steps

1. Creation of a more detailed Italian morpheme dictionary for NooJ
2. Association of each morpheme to specific keystrokes on the Stenograph key-board
3. Factorization of such associations, in order to reduce the number of keystrokes which a stenograph writer must learn to use
4. Creation of NooJ grammars for word segmentations and typo correction
5. Creation of a Python Module, as both a standalone routine and a NooJ extension
Choukran bezaf
Merci de votre attention
Thank you for your attention

Maurice Gross