


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

NOOJ 2017 CONFERENCE

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)

What is Stenotyping? - 2

- 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 alphanumeric 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."

The «Metodo Melani» - 2

- 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)

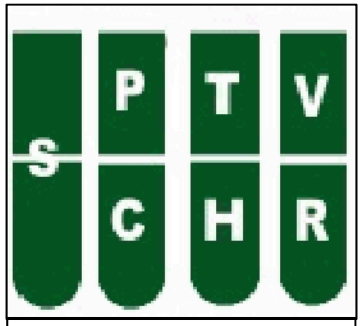
The computer input is relatively simple: there is no need for sophisticated algorithms or particularly large dictionaries of abbreviations, the computer simply recognize some shortened codes, conceived from the beginning so as not to give rise to ambiguities, and managing a dictionary of acronyms

The «Metodo Melani» - 3

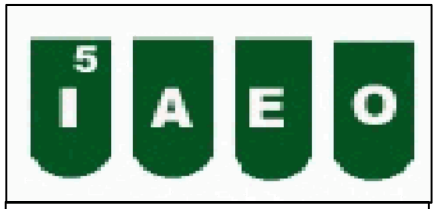
- The "Metodo Melani" **disassembles the keyboard into two basic parts: the part operated by the left hand, and the part operated by the right hand**

- 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)

The «Metodo Melani» - 5

- 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; CTPSHR = Q (six keys simultaneously pressed exactly on the splitting slot); CT = V; SH = Z**

The «Metodo Melani» - 6

- 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 Limits of the «Metodo Melani» - 1

- 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 Limits of the «Metodo Melani» - 2

- 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**

Overcoming the limits of the «Metodo Melani» - 1

- 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 (ie, the morphemes of a language) together with a production syntax (ie, the word-formation rules of a language)
- The **morphemic structures** of languages are **more stable/predictible 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**

Overcoming the limits of the «Metodo Melani» - 2

- **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** (ie groups of words sharing the same lexical morpheme) **and morpheme combinations inside words** (ie 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

Overcoming the limits of the «Metodo Melani» - 3

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

Overcoming the limits of the «Metodo Melani» - *4*

- 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**

Overcoming the limits of the «Metodo Melani» - 5

- ***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=ità
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

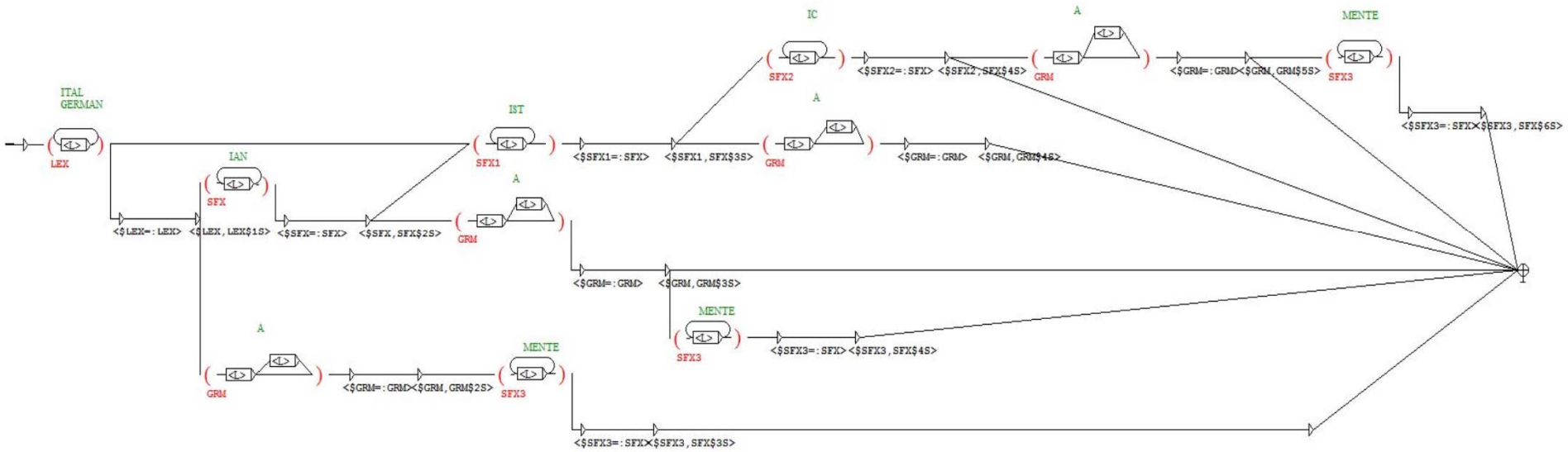
- The non-random **combination of these 27 morphemes** "produces" **161 usable Italian words**

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

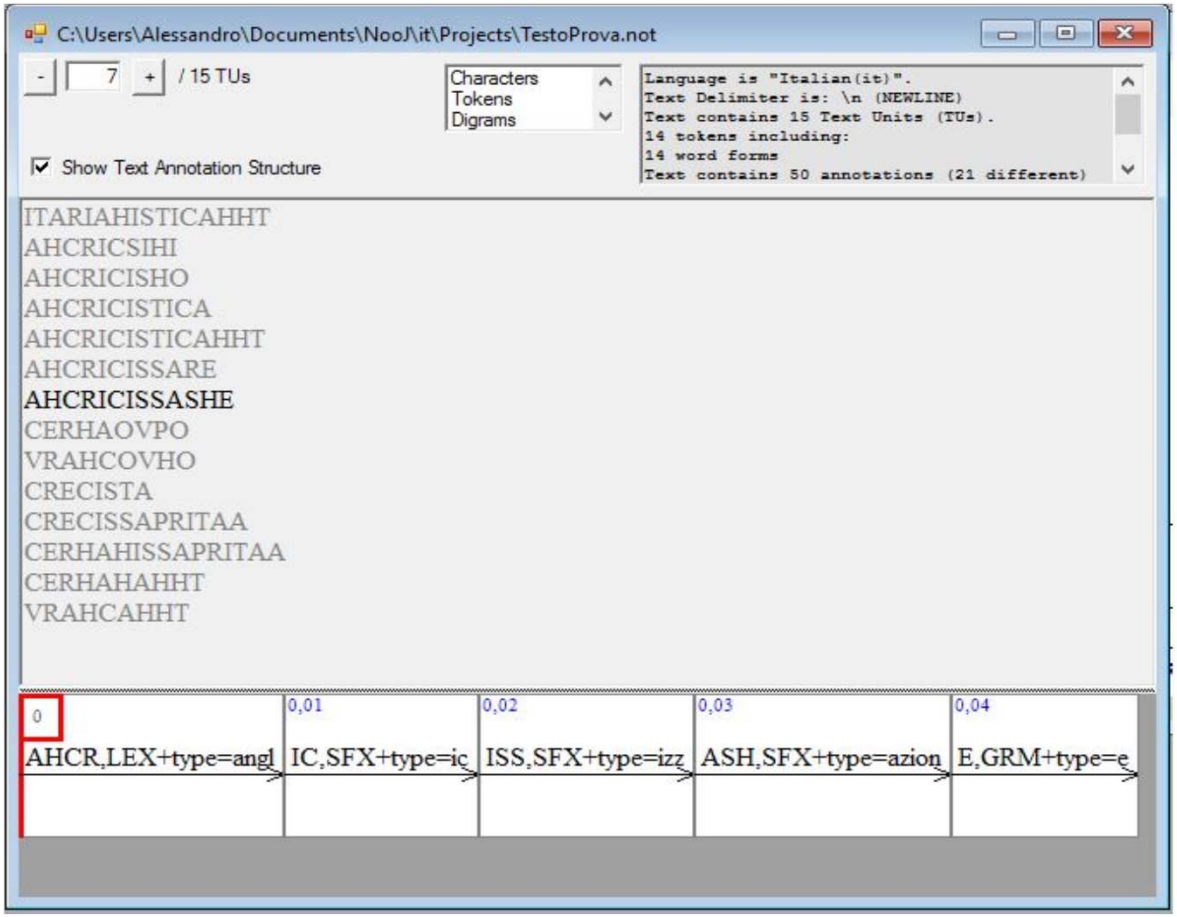
Overcoming the limits of the «Metodo Melani» - 6

- 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

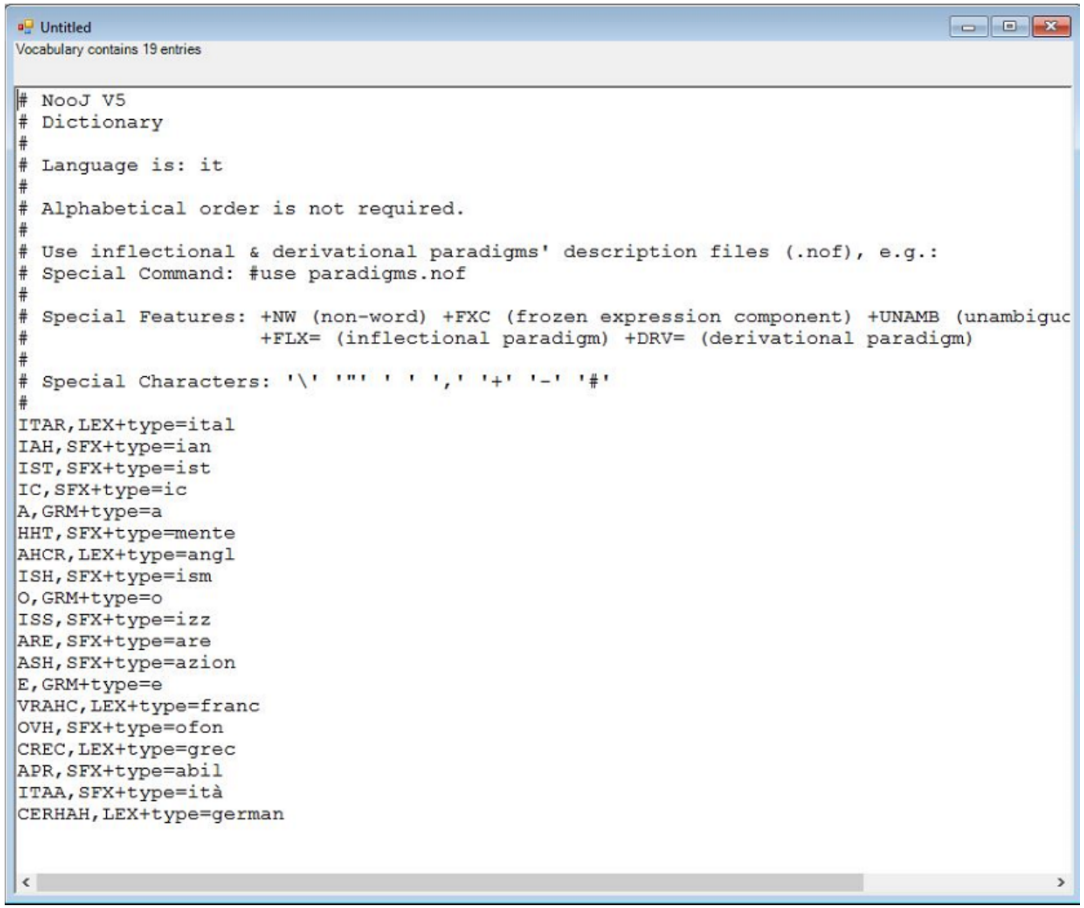


Overcoming the limits of the «Metodo Melani» - 8



Example of stenograph annotations in a text

Overcoming the limits of the «Metodo Melani» - 9



NooJ dictionary of the annotated text

Overcoming the limits of the «Metodo Melani» - 10

- **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:

```
s p t v * c t p i A
s C H r * s H R e o
    i a E o
```

- The additional successive morphemes will be written in sequence:

```
    ISS
S p t v * c t p i a
S c h r * S h r e o
    I a e o
```

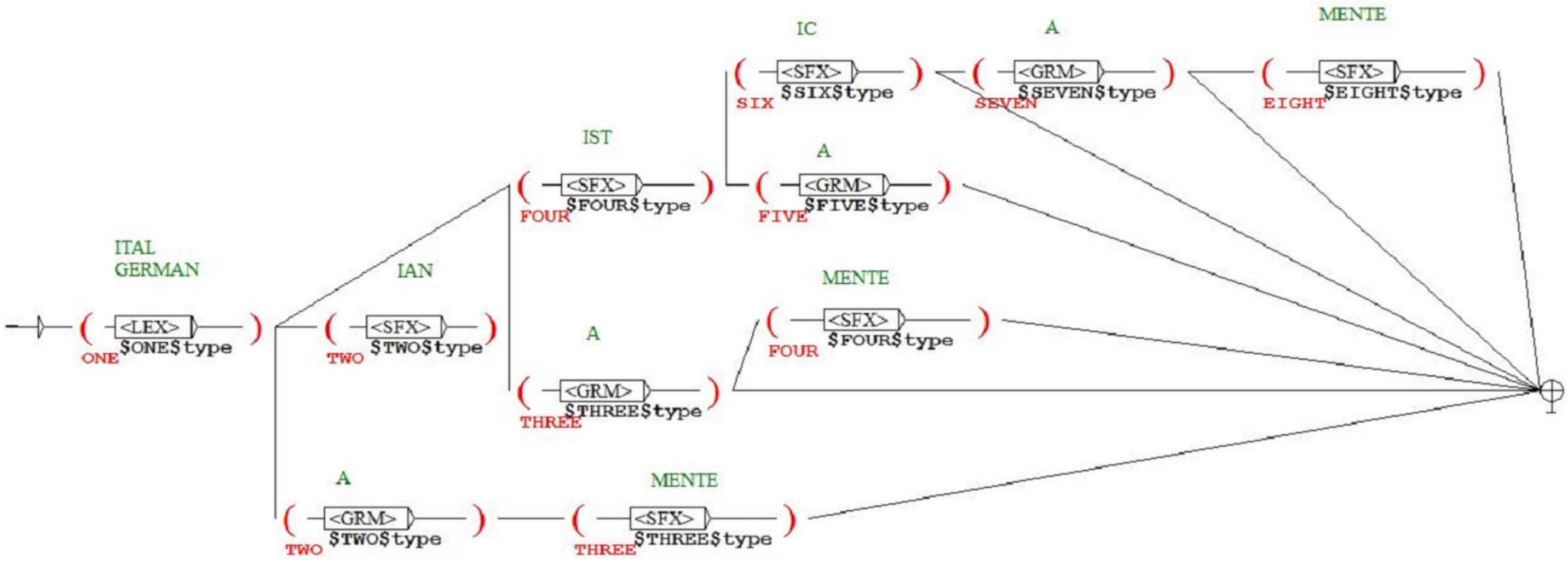
```
    APR
s p t v * c t P i a
s c h r * s h R e o
    i A e o
```

```
    ITAA
s p t v * c T p i A
s c h r * s h r e o
    I A e o
```

Overcoming the limits of the «Metodo Melani» - 11

- 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*


ITAL
GERMAN
<LEX>
ONEtype
ONE
IAN
<SFX>
TWOtype
TWO
IST
<SFX>
$FOUR$type
FOUR
A
<GRM>
$THREE$type
THREE
IC
<SFX>
SIXtype
SIX
A
<GRM>
$FIVE$type
FIVE
A
<GRM>
TWOtype
TWO
MENTE
<SFX>
$THREE$type
THREE
MENTE
<SFX>
$FOUR$type
FOUR
A
<GRM>
$SEVEN$type
SEVEN
MENTE
<SFX>
$EIGHT$type
EIGHT

Conclusions and Further Steps

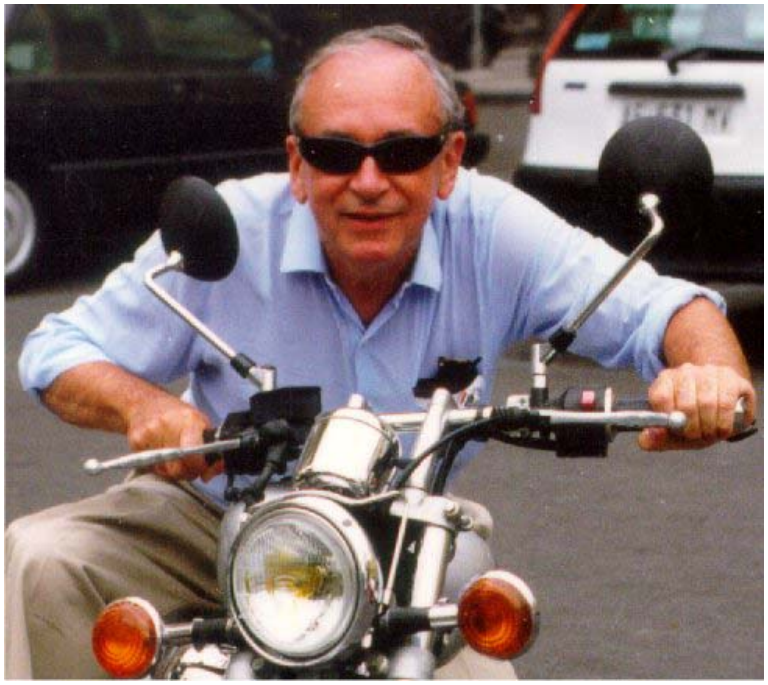
1. Creation of a more detailed Italian morpheme dictionary for NooJ
2. Association of each morpheme to specific keystrokes on the Stenograph key-board
3. Factorization of such associations, in order to reduce the number of keystrokes which a stenograph writer must learn to use
4. Creation of NooJ grammars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

نتلقاو

Choukran bezaf

Merci de votre attention

Thank you for your attention



Maurice Gross