2nd International Workshop on Phonotactics and Phonotactic Modeling - Pisa, November 21-22, 2013

An Inductive Classification of Italian Clusters

all automated! programmed in Java!

Matteo Pascoli
Università di Verona
Université de Toulouse - Le Mirail
matteo.pascoli@univr.it

words

Input data is a lexicon of Italian lemmas, with phonetic transcription made with heuristics on data from Italian online dictionaries (Treccani, Hoepli, Sabatini-Coletti) and refined by hand.

clusters

The words are divided around nuclei (vowels) to fill up a database of consonantic (including glides) clusters. Clusters are classified as initial, intervocalic, final.

sonority patterns

Each segment of the clusters is mapped to a sonority value. These sonority values are chained into a string, or sonority pattern. The clusters are then grouped according to their sonority pattern.

sonority contours

Each sonority pattern is linked with a sonority contour. Sonority contours are strings of symbols that represent the variation of sonority from a segment to the next. These symbols are: **#** for word boundary, **/** for rising sonority, **** for falling sonority, and **–** for plateau (same sonority level).

syllabification

I tried to use contours as indices for deciding where to break each cluster, that is, for the syllabification of the words. Most contours are sufficient indices, with the exception of contours 5 (\//), 8 (\///), and 16 (\\\/). For clusters belonging to these contour groups, the sonority patterns are to be examined as indices for syllabification.

- ***27.603 words***
- ***99.260 clusters***
- ***6.816 unique clusters***

the most common:

	intervocalic	*initial*	*final*
r	9758 ('ɔ.ro)	1902 (re)	22 (bar)
t	6381 (’da.ta)	873 (te)	11 (bit)
n	6265 (’u.no)	413 (ne)	23 (in)
l	4870 (’bi.le)	674 (lo)	8 (il)
k	2881 ('ɛ.ko)	2359 (ki)	7 (suk)
m	2929 (’i.mo)	1566 (mɔ)	17 (tram)
nt	3623 ('on.ta)	-	2 (sprint)
d	1608 ('ɔ.de)	1490 (di)	1 (po.la'rɔ.id)
p	1151 ('a.pe)	1603 (per)	4 (dzip)
s	901 ('fu.so)	1492 (su)	27 (bis)
v	1722 ('i.vi)	600 (vi)	-
st	1730 ('us.ta)	330 (sto)	6 (su'dɛst)
b	1164 ('rɔ.ba)	821 (bɔb)	3 (bɔb)
tt	1975 ('ɔt.to)	-	-

the less common:

	intervocalic
bstr	1 (sub'stra.to)
mbrj	1 (em.brjo'na.le)
nglj	1 ('gan.gljo)
nspj	1 (in.spje'ga.bi.le)
rflw	1 (su'pɛr.flwo)
rksj	1 (mar'ksja.no)
rstr	1 (su.per'stra.da)
bbw	1 (ab'bwɔ.no)
ddj	1 (ad'djɛ.tro)
fkj	1 (kaf'kja.no)
grj	1 (ne'grjɛ.ro)
jtj	1 (aj'tja.no)
lkw	1 (al'kwan.to)
…	

- ***sonority scale***

1	complex stops	7	voiced fricatives
2	unvoiced plosives	8	nasals
3	voiced plosives	9	laterals
4	unvoiced affricates	10	trills
5	voiced affricates	11	glides
6	unvoiced fricatives	12	vowels

- ***234 unique sonority patterns***

examples:

word	*son. pattern*	*contour*
sup'ɛrstite	12,10,6,2,12	\\\/
rejnser'ire	12,11,8,6,12	\\\/
inkwj'ɛto	12,8,2,11,11,12	\\/-/
ɲ'ɔmo	0,8,12	#//

- ***27 unique sonority contours***

contour		*tokens*	*example*	*syll. break*
1	\/	44254	d’ata, p'atʃe	V.CV
2	#/	15812	ke, ʎi	-
3	\\/	14485	‘onta, v'ispo	VC.CV
4	\-/	8536	k'ɔpto, p'aʎʎa	VC.CV
5	\//	6011	'arja, 'ɔddʒi	VC.CV, V.CCV
6	#//	3108	kr'ɛpa, tw'ɔno	-
7	\//	2142	s'ɛmpre, m'uskjo	VC.CCV
8	\///	1443	optsj'one, mell'iflwo	VC.CCV, V.CCCV
9	#\/	1441	zd'eɲɲo, sk'ɛda	-
10	\-//	937	'ɔppjo, akkl'uzo	VC.CCV
11	#\//	534	str'ega, zgw'ardo	-
12	#-/	232	zv'ago, mnem'ɔniko	-
13	\#	136	tram, il	-
14	#-//	32	sfr'edʒo, zvwotam'ento	-
15	\\#	30	nɔrd, film	-
16	\\\/	27	sup'ɛrstite, rejnser'ire	VC.CCV, VCC.CV
17	\\///	25	ind'ustrja, g'angljo	VC.CCCV
18	#///	17	trjest'ino, zmw'ɔvere	-
19	\/-/	16	el'ɔkwjo	V.CCCV
20	\/\//	11	'ɛkstra, substr'ato	VC.CCCV
21	\\\//	11	awstr'ale, s'anskrito	VC.CCCV
22	\/#	7	faks	-
23	\\/-/	4	inkwj'ɛto	VC.CCCV
24	\\-/	3	tr'ansfuga, postkontʃilj'are	VCC.CV
25	#/-/	3	kwj'ɛto	-
26	\-/-/	2	akkwjeʃʃ'ɛnte	VC.CCCV
27	\\-//	1	postprandj'ale	VCC.CCV

- ***syllable structure summary***

structure	*tokens*	*initial*		*median*		*final*		*isolate*	
CV	64014	7970	'dɛ.a	30559	e'rɔ.e	25457	'a.ja	28	ma
CVC	21936	7778	mul'la	14046	e'lɛt.ta	83	'ɛ.ros	29	bar
CCV	8907	2594	'tri.o	5123	e'brɛ.o	1181	'ɔ.ljo	9	psi
VC	4881	4562	'is.sa	311	du'ɛl.lo	6	ra'is	2	il
V	3582	1853	'ɔ.zo	1011	'ɔ.a.zi	713	'rɛ.o	5	o
CCVC	3488	2167	'skɔl.lo	1305	e'trus.ko	9	'ka.mjon	7	zlip
CCCV	408	303	'stri.a	68	u.brja.'ko.ne	37	'a.trjo	-	-
CCCVC	295	282	'stril.lo	13	ma'trjɔs.ka	-	-	-	-
CVCC	30	2	post.pran'dja.le	-	-	23	'i.noks	5	nɔrd
CCVCC	8	2	'trans.fu.ga	-	-	4	‘du.pleks	2	spɔrt
VCC	2	-	-	-	-	-	-	2	ɛks
CCCVCC	1	-	-	-	-	-	-	1	sprint