Literature and Internet Database Mining in a Study About the Word CHEMOMETRICS

Rudolf Kiralj and Márcia Miguel Castro Ferreira. *rudolf@iqm.unicamp.br, marcia@iqm.unicamp.br, http://lqta.iqm.unicamp.br*
Laboratório de Quimiometria Teórica e Aplicada (LQTA), Instituto de Química, Universidade Estadual de Campinas, Campinas SP, 13083-970, Brazil

THE OBJECTIVES OF THIS WORK

1) To use internet and literature data mining for etymological, linguistic and bibliometric/webometric study of the word CHEMOMETRICS

2) To find languages in which CHEMOMETRICS is known, and to rationalize the obtained orthographic and pronunciation forms in terms of linguistics, geography, history and science development

3) To correlate bibliometric and webometric results for CHEMOMETRICS with chemometric activities in Europe and worldwide

CHEMOMETRICS: COMPLETE ETYMOLOGY

Figure 1. CHEMOMETRICS was coined in January 1971 by Prof. Svante Wold first in Swedish (kemometri) and then in English[1]. Figure shows complete etymology of English CHEMOMETRICS and the most frequent equivalent CHEMOMETRY as well as their roots from antiquity/neolithic times to present days.

[1]Esbensen K, Geladi P (1990) J Chemometr 4:389

CHEMOMETRICS: LINGUISTIC REALITY

CHEMOMETRICS in:
48 languages, 82 orthographic forms,
127 pronunciation forms,
10 writing system – on 6 continents

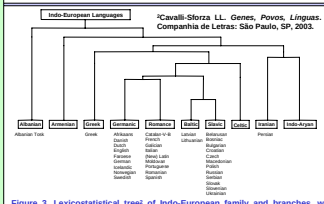
Table 1. Linguistics of the words **chemometrics** in standard national languages

Smp.[a]	Language[b]	chemometrics[c]	Latin transliteration[d]	No. Pron.Vars.[e]	Char[f]	Freq[g]
alb-1	Albanian	chemometrie	chemometrie	1	I	m
afr-2	Afrikaans	chemometrika	chemometrika	1	K	mi
alb-1	Albanian - Tosk	kemometria	kemometria	1	J	n
bas-1	Basque	kimiometria	kimiometria	1	J	n
bls-1	Belarusian	хемометрыя	khemometrýia, chemometrýia	1	J	b
bls-2	Belarusian	хемометрыка	khemometrýka, chemometrýka	1	K	b
bos-1	Bosnian	hemometrija	hemometrija	1	J	n
blg-1	Bulgarian	хемометрия	khemometrija, khemometrýja	1	J	n
cvb-1	Catalan Valencian-Balear	quimiometria	quimiometria	2	J	n
cmn-1	Chinese - Mandarin	化學計量學	hua xue jì liang xue	2	O	mi
cmn-2	Chinese - Mandarin	化学计量学	hua xue jì liang xue	2	O	mi
cmn-3	Chinese - Mandarin	化學計量學	hua xue jì liang xue	2	O	mi
cmn-4	Chinese - Mandarin	hua xue jì liang xue	hua xue jì liang xue	2	O	n
crt-1	Croatian	kemometrija	kemometrija	1	J	sh
crt-2	Croatian	kemometrika	kemometrika	1	K	hh
cze-1	Czech	chemometrie	chemometrie	1	J	m
cze-2	Czech	chemometrika	chemometrika	1	K	mi
dan-1	Danish	kemometri	kemometri	1	I	m
dan-2	Danish	kemometrik	kemometrik	1	K	mi
dut-1	Dutch	chemometrie	chemometrie	1	J	n
eng-1	English	chemometrics	chemometrics	4	K	m
eng-2	English	chemometry	chemometry	4	I	mi, mU(UK,USA)
eng-3	English	chemiometrics	chemiometrics	4	K	n
eng-4	English	chemiometry	chemiometry	4	I	n
eng-5	English	chemimetrics	chemimetrics	4	K	n
eng-6	English	chemimetry	chemimetry	4	I	n
est-1	Estonian	kemeometria	kemeometria	1	J	n
far-1	Faroese	kemeometri	kemeometri	1	I	n
fin-1	Finnish	kemometria	kemometria	1	J	n
fre-1	French	chimiométrie	chimiometrie	2	I	mm
fre-2	French	chimiométrie	chimiometrie	1	I	mi
fre-3	French	chemiométrie	chemiometrie	1	I	n
fre-4	French	chémiométrie	chemiometrie	1	I	n
gal-1	Galician	quimiometria	quimiometria	1	J	n
ger-1	German	chemometrie	chemometrie	3	I	mm
ger-2	German	chemometrik	chemometrik	3	K	mi
grk-1	Greek	χημειομετρία	khimeiometría, khimiometría	1	J	n
hng-1	Hungarian	kemometria	kemometria	1	J	n
icl-1	Icelandic	chemometrics	chemometrics	1	K	n
ind-1	Indonesian	kemometri	kemometri	1	I	h
ind-2	Indonesian	kemometrik	kemometrik	1	K	hh
ind-3	Indonesian	kemometrika	kemometrika	1	K	mi
itl-1	Italian	chemiometria	chemiometria	1	J	mm
itl-2	Italian	chimiometria	chimiometria	1	J	n
jpn-1	Japanese	ケモメトリックス	kemometorikkusu	1	O	mm
jpn-2	Japanese	計量化学	keiryou kagaku	1	O	mi
jpn-3	Japanese	けりょうかがく	keiryou kagaku	1	O	n
jpn-4	Japanese	kemometori	kemometori	1	I	n
kor-1	Korean	계량분석화학	yae ryang pun seok hoe hak, yay ryang pun seok hoe hak	1	O	n
lat-1	Latin	chemometria	chemometria	1	J	m
lat-2	Latin	chimiometria	chimiometria	1	J	mi
lav-1	Latvian	hemometrija	hemometrija	1	J	h
lav-2	Latvian	kemometrija	kemometrija	1	J	hh
lav-3	Latvian	kemometrika	kemometrika	1	K	mi
lit-1	Lithuanian	chemometrija	chemometrija	1	J	n
mcd-1	Macedonian	хемометрија	hemometrija	1	J	n
mal-1	Malay	kimometrik	kimometrik	1	K	m
mal-2	Malay	kemometrik	kemometrik	1	K	mi
mld-1	Moldovan	hemometrica	hemometrica	1	K	n
nrw-1	Norwegian	kjemometri	kjemometri	1	I	n
per-1	Persian	شیمی‌سنجی	shimi'sa' ámjiyh', ilmiyn' sinjiyh'	1	O	n
pol-1	Polish	chemometria	chemometria	1	J	n
prt-1	Portuguese	quimiometria	quimiometria	2	J	n
rum-1	Romanian	chemometrie	chemometrie	1	J	mm
rum-2	Romanian	chemometria	chemometria	1	J	mi
rus-1	Russian	хемометрика	khemometrika	1	K	mm
rus-2	Russian	хемометрия	khemometrija, khemometrýja	1	J	mi
srb-1	Serbian	хемометрија, hemometrija	hemometrija	1	J	m
srb-2	Serbian	хемометрика, hemometrika	hemometrika	1	K	mi
slk-1	Slovak	chemometria	chemometria	1	J	n
slv-1	Slovenian	kemometrija	kemometrija	1	J	mm
slv-2	Slovenian	kemometrika	kemometrika	1	K	n
spn-1	Spanish	quimiometria	quimiometria	2	J	mm
spn-2	Spanish	quimiometría	quimiometría	2	J	mi
swe-1	Swedish	kemometrik	kemometrik	1	K	n
tha-1	Thai	เคมีมาตรวิทยา	kam ra ra wi tha khong khemi	1	O	m
tha-2	Thai	เคโมเมตริกส์	khemometrikis	1	K	mi
trk-1	Turkish	kemometri	kemometri	1	I	n
ukr-1	Ukrainian	хемометрія	khemometrýia, khemometrýja	1	J	m
ukr-2	Ukrainian	хемометрика	khemometrika	1	K	mi
vtn-1	Vietnamese	thuốc đo hoá học	thuốc đo hoá học	1	O	n

[a]Smp. – sample – a written form of the word **chemometrics** in a particular language.
[b]Language – language name as recognized in linguistic, historical, cultural and political terms.
[c]**chemometrics** – a written form of the word **chemometrics** in a particular language.
[d]Latin transliteration – the most common way(s) of Latin transliteration for forms written in non-Latin writing systems.
[e]No. Pron. Vars. – number of pronunciation variants, taking into account No. standards of a particular language.
[f]Char – the pronunciation end-types: I (the end with long or short [I]-type sound), J (end with semi-vowel [j] or vowel combinations ne in [i-e], [ini], [i-e], [jni], etc.), K (end with [k]-type sound as [ks], [ks], or [k]), and O (other end-forms in morphologically different words – in Asian languages).
[g]Freq. – relative frequency of the samples as found in the internet resources: n – unique (100%), m – almost unique (>99%), mm – absolute majority (90-99%), m – major (60 – 90%), sh – above half (53-60%), h – around half (47 – 53%), hh – below half (40-47%), mi – minor (10-40%), mi – minority (0.5-10%), and n – negligible (<0.5%).

CHEMOMETRICS: LINGUISTIC, GEOGRAPHIC AND HISTORICAL TIES

Figure 2. Official science and higher education languages in Europe (left) and Asia (right) and their linguistic classification and geographic distribution. Underlined are languages for which CHEMOMETRICS is known in Asia. The languages in Asia marked with (*) have at least one form for CHEMOMETRICS consisting only of CHEMO- and -METRICS. The influence of Indo-European languages in Asia is obvious.



²Cavalli-Sforza LL. *Genes, Povos, Línguas*. Companhia de Letras: São Paulo, SP, 2003.

Figure 3. Lexicostatistical tree² of Indo-European family and branches, with particular languages in which CHEMOMETRICS is known.

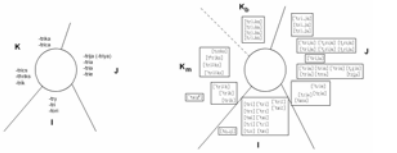


Figure 4. CHEMOMETRICS (-ICS) versus CHEMOMETRY (-Y) in all languages? Putative classification of orthographic (left) and pronunciation (right) end forms of CHEMOMETRICS exhibits three main groups (K, I, J) and not only two as many chemometricians would expect (K and I, equivalent to -ICS and -Y, respectively). This analysis is based on the data from extended Table 1.

Figure 5. Geographic distribution of CHEMOMETRICS orthography & pronunciation in Europe. Left: CHEM- particle. Middle: -MO- (-MIO) particle. Right: -TRIX particle.

Figure 6. Geographic frequency distribution of CHEMOMETRICS in Europe. Left: Predominant orthographic forms. Middle: CHEMOMETRICS versus CHEMOMEMTRY in WOS publications. Right: No. Google hits for CHEMOMETRICS in national languages.

EXPLORATORY ANALYSIS & DATA MINING

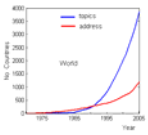


Figure 7. Growth of No. WOS publications with CHEMOMETR* in topics and address. Exponential-like increase is visible for the both growths.

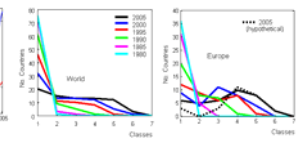


Figure 8. Distribution function of No. WOS publications with CHEMOMETR* in topics. Classes belong to log units of this number: 1 (0-0.5 units), 2 (0.5-1), 3 (1-1.5), 4 (1.5-2), 5 (2-2.5), 6 (2.5-3), and 7 (3-3.5). Hypothetical Europe means without disruption of USSR, Czechoslovakia and Yugoslavia. The tendency of normal curve formation is visible, especially in Europe. Eastern Europe political changes slow down this trend.

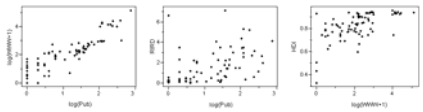


Figure 9. Five descriptors for Europe and the world were generated: log(Pub) – log of No. WOS publications with CHEMOMETR* in topics; 1stpub – the corresponding first publication date; log(WWW+1) – log form of the No. Google hits for English CHEMOMETRICS (WWW); HDI – 2002 human development index; RIRD – researchers in research and development index for 1990-2001. The correlograms are for Europe, similar to those for the world (not shown).

Figure 10. PCA scores plot for Europe, using the five autoscaled descriptors. The countries are marked with internet country domain names. The highest organizational level of chemometrics is marked by diamonds: chemometric society (red), chemometric laboratories (blue), and other (green). The countries with WOS publications with CHEMOMETR* in address are marked by empty red squares. With exception of CH, TR and IE, higher chemometric activity is concentrated in the region of red/blue diamongs and red squares. Analogue and more complete situation is observed for the world (not shown), what means that chemometrics is closely related to scientific and technological development.

ACKNOWLEDGEMENT: FAPESP