Exploring Register Variation in Turkish Web Corpus

Selcen Erten
University of Turku, Finland
seerte@utu.fi

Abstract

In linguistics, web registers are language varieties occurring on the web such as *news reports* and *editorials*. Most of the previous web register research has been done for Indo-European languages. Additionally, previous research has mainly focused on the restricted corpora with pre-determined registers. This article describes Turkish web registers on the web. The data is Turkish web register corpus which consists of 2601 web texts. A taxonomy was adapted to register label these texts. The manual annotations of the texts were done with the adapted taxonomy, and the registers were defined accordingly. Text dispersion keyword analysis was used to generate the keywords of the registers and examine the basic linguistic characteristics of them. The results display the web registers existing for Turkish, and the linguistic characteristics associated with the *news report* and *editorial* registers.

Keywords: Turkish web registers, manual annotation, text dispersion keyword analysis.

1. Introduction

In linguistics, registers are language varieties written in a particular situation with pervasive linguistic features that serve important functions within that situation of use. (Biber, 1988; Biber & Conrad, 2019). Considering that the web is possibly the first source one resorts to when seeking information, it is important to understand registers on the web. Registers occurring on the web are called web registers. Some examples of them are *news reports and editorials*. Understanding web registers is crucial to be able to distinguish, e.g., facts from opinions and advertisements from informative texts.

Register studies have a relatively long history in linguistics (Biber, 1988). However, most of the previous research has been restricted to English and other Indo-European languages (Biber & Finegan, 1994; Conrad & Biber, 2001; Asencion-Delaney & Collentine, 2011; Berber-Sardinha et al., 2014; see, however, Kim & Biber, 1994; Jang, 1998; Ravid & Berman, 2009; Aksan & Aksan, 2015). Additionally, previous research has mainly focused on carefully curated restricted corpora, where the documents have been selected manually from established sources featuring pre-determined registers. This has led to a situation where registers are typically examined in discrete classes where texts are very good examples of their categories. The web, on the other hand, offers a very different perspective to register variation by including a wide and sometimes noisy range of documents (Biber & Egbert, 2018). There are no gatekeepers to ensure that documents follow the guiding principles of specific registers on the web. Further, not all documents have a single register or any register at all (Santini, 2007; Egbert et al., 2015). By taking the full variation into account, a much more complete understanding of web register variation can be gained than what the current studies based on restricted samples can offer.

2. Present Study

In the current study, web registers in Turkish are defined by using a web register taxonomy adapted from English (Egbert et al., 2015) and Finnish (Laippala et al., 2019) to Turkish. Although the pioneering register studies have been done in Indo-European languages, English being the most studied, understanding language use on the internet will be restricted if the examination of registers is limited to a certain group of languages. For this reason, culturally and linguistically different languages need to be studied so that more understanding of web registers can be acquired not only for language research but also for the applications of web registers in the areas of media literacy and inter-cultural communication. Turkish is culturally and linguistically a very different language than the commonly studied Indo-European languages. Considering that people are regularly surrounded by media, and there is little incentive to employ an 'off' switch (Butler, 2020), it is vital that people know how to be able to judge what is useful and misleading information and when media can be trusted (Livingstone, 2018). Further, all cultures use language for different communicative purposes in different situations. Registers are based on pervasive patterns of linguistic variation across such situations (Biber & Conrad, 2019; Biber et al., 2020). Understanding Turkish web registers will help uncover misunderstandings and failures in inter-cultural communication.

In this study, registers displaying the features of *news reports* and *editorials* are specifically examined with text dispersion keyword analysis to see their basic linguistic characteristics. *News reports* are texts typically written by professionals to report on recent events while *editorials* are texts typically written by professionals on a news-related topic with a purpose to persuade the reader about opiniated points.

In the light of the aims of defining Turkish registers on the web and examining the linguistic characteristics of them, the following research questions are answered:

1. Which web registers exist for Turkish in terms of the defined categories by Egbert et al. (2015) and Laippala et al. (2019)?
2. What are the basic linguistic characteristics of *news report* and *editorial* registers?

3. Data and Methodology

3.1. Data

The data in the study is Turkish web register corpus. The corpus targets the full Turkish speaking web. It is based on a random sample of the web, originally computationally

collected by Common Crawl (commoncrawl.org) and cleaned and pre-processed within *Massively Multilingual Modelling of Registers in Web-scale Corpora* project run by TurkuNLP team at the University of Turku, Finland. Altogether, the corpus consists of 3767 unique web texts covering various domains in Turkish.

Regarding the ethical issues in the phase of collecting data for the Turkish web register corpus, the guidelines published by the Finnish National Board on Research Integrity TENK and the Turkish Council of Higher Education on Scientific Research Directive were followed. Upon the completion of manual annotations, it was assured that there are no personal elements in the data collected and used for this research. The data is openly and freely accessible on the web.

3.2. Methodology

The taxonomy used to register-label the web texts were adapted from Egbert et al. (2015) and Laippala et al. (2019) to Turkish and its specificities. The benefit of using this taxonomy is that it allows to annotate registers in a wide range of different types of documents with no dependence on pre-determined categories. Manual annotations of each text in the corpus were completed with the adapted taxonomy on the annotation tool Prodigy. On Prodigy, the texts were first accepted or rejected. Accept means that the text was put in the data, and it was register-labelled. The document was rejected in the situations such as where the text consisted of only short list of items, the sentences did not form a coherent text, the amount of coherent text was very small compared to the junk text or the text was not in the target language (Biber & Egbert, 2018). In the data, around 35% of the texts were rejected for one or more of these reasons. Some of the accepted documents were annotated as hybrids, which means that they were given two labels or more although it was typically two. This occurred when a text featured characteristics of more than one register such as a marketing text followed by reviews (description with intent to sell + review). Compared to the single-category registers where each text had only one register, the hybrid texts were a few. Hybrid registers were not included in this study, and only 2601 accepted, single-category registers were examined.

In addition to the manual annotations of the Turkish web texts, keyword analysis method was used to examine the basic linguistic characteristics of *news reports* and *editorials*. The concept of keyness in text have been discussed in various ways (Scott, 1997; Bondi, 2010; Culpeper & Demmen, 2015), yet there are two fundamental approaches in corpus linguistic methods which determines the keyness in frequency (Scott & Tribble, 2006) and in dispersion (Egbert & Biber, 2019). In this study, text dispersion keyword analysis was used, as it is seen as the most suitable method for register studies with large corpora containing many texts. Text dispersion keyness uses the text, rather than the corpus, as the unit of observation. It is based on a word's dispersion across the texts of a corpus rather than its overall frequency in the corpus (ibid). This means that text dispersion keyness disregards word frequency entirely but generates keyword lists based on word dispersion across texts. Log-likelihood is used as it estimates probabilities more accurately even when the counts are low, and because the dispersion of the words across texts tend to follow a Zipfian distribution. The requirements for text dispersion keyness are many texts in target and reference corpora as well as a special program. In frequency-based keyness, the most frequent words are general high-frequency words which are not particularly distinctive to the target corpus. The text dispersion method, on the other hand, identifies words which are much more strongly related to the target corpus than the reference corpus. In the current study, both the target and reference corpora were generated from the web text data. If, for example, texts of news reports were the target corpus, the reference corpus was all the other texts belonging to various registers minus news reports. As for the special program to acquire the text dispersion values, Python codes were utilised for the purpose.

4. Results

4.1. Web registers of Turkish

Based on the taxonomy adapted to Turkish, 9 main register categories and sub-registers falling under them were identified.

Below, the web registers defined for Turkish are displayed without the distinction between main or sub-registers. The total number of texts and number of characters for each register are also as in the following:

Register	Number of texts	Number of char.
Description with intent to sell	645	1,920,852
News report	556	1,502,494
Machine translated	329	1,819,394
Other-informational description	224	954,392
Description of a thing or person	124	524,602
Legal terms	105	593,928
Editorial	93	774,258
Review	66	240,271
Opinion blog	58	377,527
Narrative blog	52	325,364
Interactive discussion	50	595,558
Advice	46	193,699
Recipe	40	81,555
Other-informational persuasion	40	103,980
Sports report	30	66,565
Religious blog	29	323,391
Other-spoken	20	59,908
Other-how-to or instruction	22	62,427
Encyclopaedia article	18	93,131
Other-opinion	16	132,042
Lyrical	16	39,810
Interview	12	107,967
FAQ	6	27,030
Research article	4	19,649
Total	2601	10,939,794

Table 1: Registers identified in Turkish web register corpus with their numbers of texts and characters.

As seen, there are *other* categories among the registers identified for Turkish. *Other* means that the text fell under one of the main categories, but it could not completely be annotated as one of the sub-categories of the main category. Although this study does not focus on the *other* categories,

they might in fact show language and culture-specific features.

4.2. Linguistic characteristics of news reports and editorials

Egbert & Biber (2019: 87) state that the top 100 keywords suffice to show the strengths of the text dispersion keywords.

The top 100 keywords for news reports and their values of keyness are displayed below:

	Keyword	Translation	Keyness
1	dedi	s/he said	295,277
2	başkanı	chairman of	289,815
3	konuştu	s/he spoke	202,403
4	söyledi	s/he said	167,021
5	başkan	chairman	142,405
6	etti	s/he did	138,902
7	kullandı	s/he used	128,390
8	belediye	municipality	127,679
9	ifadelerini	expressions of	112,388
10	edildi	it was done	109,101
11	belirtti	s/he indicated	106,923
12	belirten	… who indicated	104,918
13	bulundu	it was found	103,052
14	kaydetti	s/he noted	92,111
15	açıklamada	in the statement	89,773
16	belediyesi	municipality of	88,350
17	bin	thousand	84,626
18	belirterek	by indicating	83,929
19	koronavirüs	coronavirus	82,141
20	verdi	s/he gave	73,258
21	yapıldı	it was done	68,539
22	bakanı	minister of	67,920
23	katıldı	s/he participated	65,634
24	müdürü	director of	62,220
25	ifade	expression	60,086
26	kovid	covid	59,936
27	ilçe	county	58,616
28	haber	news	57,432
29	yaptığı	…which s/he did	57,014
30	alındı	it was taken	55,330
31	chp	chp (republican party)	54,503
32	sözlerine	to the statements of	54,251
33	il	province	53,835
34	ardından	afterward	53,394
35	büyükşehir	metropolis	52,103
36	ekipleri	teams of	51,055
37	itfaiye	fire department	50,234
38	ilçesinde	in the county of	49,755
39	mustafa	mustafa	48,647
40	onaylanmamaktadır	it is not (being) approved	47,076
41	olay	incident	46,793
42	dile	to the tongue	46,544
43	öğrenildi	it was learnt	46,308
44	müdürlüğü	directorship of	46,241
45	mehmet	mehmet	46,197
46	yardımcısı	vice of	45,138
47	heyeti	board of	43,955
48	harflerle	with the letters	42,862
49	bildirildi	it was informed	41,520
50	'	'	41,228
51	şunları	those	41,168
52	belirtildi	it was stated	40,975
53	edinilen	…which was acquired	40,080
54	dr	dr (doctor)	39,899
55	açıkladı	s/he explained	38,371
56	milyon	million	38,370
57	verildi	it was given	37,999
58	hesabından	from the account of	37,422
59	aa	aa (anatolian agency)	36,433
60	parti	party	35,683
61	basın	press	35,259
62	soruşturma	investigation	35,121
63	vurgulayan	…who underlined	34,899
64	konuşan	…who spoke	34,779
65	vurguladı	s/he underlined	34,761
66	19	19	34,721
67	polis	police	34,115
68	kullanılmayan	…which was/is not used	34,115
69	ekiplerinin	of the teams of	34,077
70	muhabirine	to the journalist of	34,077
71	vatandaşlar	citizens	33,988
72	ilişkin	related	33,782
73	söyleyen	…who told	33,602
74	milletvekili	congressman	33,444
75	katıldığı	…which s/he participated	33,444
76	inşallah	God willing	33,276
77	yaşındaki	in the age of	33,276
78	vali	governor	32,831
79	devam	continuation	32,804
80	gerçekleştirildi	it was fulfilled	32,630
81	gözaltına	to the custody	32,321
82	tedbirleri	precautions of	31,850
83	değinen	…who mentioned	31,091
84	salonunda	in the hall of	31,068
85	salgını	epidemic of	31,068
86	toplantısında	in the meeting of	31,068
87	yüzde	per cent	29,993
88	başlatıldı	it was started	29,968
89	açıklamalarda	in the statements	29,771
90	saatlerinde	in the time of	29,732
91	kaydeden	…who noted	29,514
92	olayla	with the incident	29,511
93	jandarma	gendarme	29,404
94	yaralı	injured	29,404
95	recep	recep	28,778
96	düzenlenen	…which was organized	28,344
97	içişleri	internal affairs	28,150
98	yaralandı	s/he was injured	28,118
99	tarım	agriculture	27,618
100	önümüzde	ahead of us	27,617

Table 2: Top 100 text dispersion keywords of news reports and their values of keyness.

Closer inspection of the table shows that 53 keywords emerge as nouns, 33 keywords as verbs and 14 as *other*.

Among the nouns, most nouns are administration words such as *chairman, municipality, minister, director, board, citizen, congressman* and *governor.* There are other nouns falling under the themes of disaster (*covid, coronavirus, fire department)*, legality (*investigation, police, custody and gendarme)*, journalism (*aa, journalist, press, news)* and communication (*expression, statement, utterance).*

Among the verbs, 23 of them are finite verb forms (predicates) while 13 of them are non-finite. With one exception, all predicates have *past tense + 3rd person singular* pattern, which seems to be the pattern for news reports that report what happened. In addition to *past tense + 3rd person singular* pattern, half of the predicates have *passive voice,* which also emerges as a pattern in news reports where the action is important. Passive voice also emerges in non-finite verb forms of the keywords of news reports. There are three non-finite verb forms in Turkish, which are verbal nouns, participles and converbs (Göksel & Kerslake, 2005). In the keyness of news reports, all non-finite verb forms, with one exception, were found to be as participles: non-finite verb forms of relative clauses formed with *who* and *which*. When both predicates and non-finite verb forms are considered together, it is seen that there are many communication verbs such as *say, tell, note, underline, inform, state* and *explain* used in news reports.

When it comes to editorials, the striking thing featuring for the keyness of editorial texts is that seven different part-of-speech classes and *other*-category were identified:

42 nouns,	5 adverbials,
13 discourse connectives,	4 postpositions,
13 adjectives,	4 pronouns,
7 verbs,	12 *other*-category words.

Top 100 text dispersion keywords for editorials and their values of keyness are as in the following:

	Keyword	Translation	Keyness
1	iktidar	rulership	73,517
2	ama	but	50,719
3	meselesi	matter of	48,418
4	bile	even	47,603
5	üstelik	what's more	46,382
6	karşı	against	43,233
7	yok	nonexistent, no	40,817
8	devlet	state	39,631
9	siyasi	political	39,364
10	ne	what	38,337
11	işte	"işte" (discourse connective)	37,848
12	demokratik	democratic	37,693
13	değil	not	37,681
14	asıl	actual	36,939
15	mi	"mi"	35,729
16	aslında	in fact	35,048
17	çıkarları	benefits of	35,007
18	kapitalizmin	of capitalizm	34,625
19	halk	public	34,458
20	devletin	of the state	34,180
21	erdoğan	erdoğan	34,018
22	erdoğanı	erdoğan-accusative	33,580
23	düpedüz	sheerly	33,579
24	devrimci	revolutionary	32,875
25	dedikleri	..which they say	32,874
26	oysa	though	32,623
27	önünde	in front of	32,418
28	yana	sideways	31,565
29	mı	"mı"	31,484
30	o	she/he/it	31,218
31	çünkü	because	31,118
32	akp	akp (ruling party)	30,647
33	dı	"dı"	30,238
34	peki	well then	30,116
35	artık	now/anymore	30,114
36	çıkmış	ensued/out of joint	30,061
37	türkiyeyi	Turkey-accusative	29,706
38	öyle	as such	29,207
39	biçimi	way of	28,643
40	cumhurbaşkanının	the president's	28,244
41	daha	more, yet	28,036
42	başkanlık	presidency	27,657
43	zaten	already, anyway	27,122
44	parti	party	26,912
45	propaganda	propaganda	26,864
46	ağustosta	in August	26,863
47	protesto	protest	26,851
48	sözde	so-called	26,850
49	ortaya	into the pot	26,848
50	hedef	target	26,840
51	ettiği	...which s/he/it did/does	26,833
52	azından	least	26,728
53	var	there is/are	26,436
54	diye	called, in case	26,124
55	vardı	there was/were	26,036
56	ona	to him/her/it	26,036
57	gibi	as, like	26,032
58	demokrasi	democracy	25,879
59	erdoğanın	erdogan's	25,878
60	tarihsel	historical	25,878
61	değişen	changing	25,748
62	muhalif	opponent	25,720
63	siyasal	political	25,719
64	vardır	there is/must be	25,705
65	-dır	"-dır"	25,575
66	cumhurbaşkanı	president of	25,568
67	iki	two	25,502
68	tümü	all-accusative	25,492
69	ülkenin	of the country	25,470
70	böyle	like this	25,448
71	yol	way	25,447
72	müslüman	muslim	25,280
73	şöyle	as such	24,882
74	diyerek	by saying	24,847
75	neden	why, reason	24,709
76	tur	round	24,349
77	gerçi	actually	24,309
78	dini	religion-accusative	24,298
79	çıkar	benefit	24,153
80	yani	namely	24,149
81	bizim	our	24,126
82	savaş	war	23,944
83	hiç	nothing, none	23,881
84	ya	"ya"(discourse connective)	23,395
85	diyor	s/he says	23,210
86	kesimleri	parts of	23,207
87	yıl	year	23,102
88	gazeteci	journalist	22,906
89	yaşanan	…which is/was encountered	22,872
90	toplumsal	societal	22,856
91	vatan	homeland	22,854
92	işgal	invasion	22,854
93	iktidarı	rulership of	22,847
94	işin	of the matter	22,843
95	toplum	society	22,506
96	kendini	oneself-accusative	22,391
97	insan	human	22,155
98	mesele	matter	22,035
99	batı	west	21,966
100	esad	esad	21,931

Table 3: Top 100 text dispersion keywords for editorials and their values of keyness.

The number of frequencies of the words is not in the scope of text dispersion keyness or of the current study. Nevertheless, the results showed that there is a variety of part-of-speech classes for editorial texts especially compared to news reports.

Among the nouns, most nouns are governance-related and political words such as *rulership, state, presidency* and *democracy.* The rest of the words falls under the themes of strategy (political strategy word *propaganda*, military strategy word *invasion* and economic strategy word *capitalism*), society (*society, human, public*), direction (*middle, west, target*), and belief (*religion, Muslim*).

The adjectives were found to be the words showing clear opinion and stance of the editorial writers such as *political, democratic, revolutionary* and *so-called.*

Most of the conjunctions and discourse connectives such as *but, in fact, whereas* and *actually* have adversative function. Further, there are some that have the examples of additive (*even, what's more*), causal (*because*), corroborative (*in any case*), expansive (*in other words*) and organizational (*"işte"*) functions.

The linguistic features of news reports and editorials show that news reports have only two part-of-speech classes and the *other*-category while editorials have seven part-of-speech classes and the *other*-category. Out of these categories, while past tense and passive voice are very regular in the predicates used in news reports, none of them appear for the predicates of editorials. The use of adjectives shows differences in news and editorials, as well. For

editorials, the adjective use covers 12 % of the keywords, yet for news, it is only 1 %. Another distinctive feature between news and editorial texts is the use of conjunctions and discourse connectives. Out of the top 100 keywords, 13 are conjunctions and discourse connectives in editorial texts while in news reports, it is none.

Editorial writers seem to use a variety of language to persuade the reader while news report writers report the recent events with less variety of language. News reports in Turkish seem to report on the recent happenings with the *-DI* perfective suffix rather than the *-mIş* evidential/perfective suffix. It might be a linguistic feature specific to news reports with a purpose to look more factual and updated, as the *-DI* perfective suffix in Turkish is used when the person witnessed the happenings and *-mIş* is used when the person learnt it through outer sources. In editorials, any of the past tenses do not seem to be a typical use, but adjectives seem to exist unlike in news reports. This might indicate that the use of adjectives is useful for editorial writers to strengthen their personal opinions based on reality while news reporters rather focus on reporting what happened. The conjunctions and discourse connectives usage with various functions in editorials also indicate that editorial writers' informational persuasion is provided with tailored choices of conjunctions and discourse connectives. For news reporters, this does not seem to be the case for a purpose of persuasion. It might be possible to state that they rather aim to look more factual and updated, and they seem to do it with the *-DI* perfective suffix in Turkish.

5. Conclusion

Understanding different language varieties on the web is important in an era when most of the information one needs is acquired from the web. Having the data provided from the Turkish web, Turkish web register corpus has text samples on a large variety of registers with 24 different categories.

News reports and editorials are two typical web registers which have many samples on the Turkish web but have distinctive linguistic characteristics. Understanding the differences between these registers on the web is crucial to be able to differentiate facts from opinions. While acquiring information from the web, it is important to understand which features of the language are preferred and what the purpose of the texts are so that media literacy as well as inter-cultural communication are successfully accomplished.

6. References

Asención-Delaney, Y., & Collentine, J. (2011). A Multidimensional analysis of a written L2 Spanish corpus. *Applied Linguistics*, 32, pp. 299–322.

Aksan, Y. & Aksan, M. (2015). Multi-word units in informative and imaginative domains. In *The 16th International Conference of Turkish Linguistics*. Ankara: Middle East Technical University.

Berber-Sardinha, T.; Kauffman, C. & Acunzo, C. M. (2014). Dimensions of register variation in Brazilian Portugese. In T. Berber-Sardinha & M. Veirano-Pinto (Eds.), *Multi-dimensional analysis, 25 years on: A tribute to Dougles Biber*, Philadelphia: John Benjamins, pp. 35-80.

Biber, D. (1988). *Variation Across Speech and Writing*. The UK: Cambridge University Press.

Biber, D. & Finegan, E. (1994). Multi-dimensional analyses of author`s styles: Some case studies from the eighteenth century. In D. Ross and D. Bring (Eds.), *Research in Humanities Computing*. Oxford: University Press, pp. 3-17.

Biber, D. & Egbert, J. (2018). *Register Variation Online*. Cambridge: Cambridge University Press.

Biber, D. & Conrad. S. (2019). *Register, Genre, and Style*. 2nd ed. the UK: Cambridge University Press.

Biber, D., Egbert, J. & Keller, D. (2020). Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, 16 (3), pp. 581-616.

Bondi, M. (2010). Perspectives on keywords and keyness. In M. Bondi & M. Scott (Eds.), *Keyness in Texts*. John Benjamins, pp. 1-20.

Butler, A. T. (2020). *Educating Media Literacy: The Need for Critical Media Literacy in Teacher Education.* Leiden and Boston: Brill Sense.

Conrad, S. & Biber, D. (2001). *Variation in English: Multi-dimensional Studies*. Eastbourne: Pearson Education.

Culpeper, J., & Demmen, J. (2015). Keywords. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of Corpus Linguistics*. Cambridge University Press, pp. 90-105.

Egbert, J.; Biber, D. & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66 (9), pp. 1817-1831.

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora,* 14 (1), pp. 77–104.

Göksel, A. & Kerslake, C. (2005). *Turkish: A Comprehensive Grammar.* London: Routledge.

Jang, S. C. (1998). *Dimensions of spoken and written Taiwanese: A corpus-based register study.* Unpublished doctoral dissertation. University of Hawaii, Manoa.

Kim, Y. J., Biber, D. (1994). A corpus-based analysis of register variation in Korean. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register*. Oxford: Oxford University Press, pp. 157-181.

Laippala, V.; Kyllönen, R.; Egbert, J.; Biber, D.; & Pyysalo, S. (2019). Toward multilingual identification of online registers. In *Proceedings of the Northern European Association for Language Technology*. Turku, Finland, pp. 292-298.

Livingstone, S. (2018, July 27). Media literacy-Everyone's favourite solution to the problems of regulation. Parenting for a Digital Future. Retrieved from https://blogs.lse.ac.uk/mediapolicyproject/2018/05/08/media-literacy-everyones-favourite-solution-to-the-problems-of-regulation/

Ravid, D. & Berman, R. (2009). Developing linguistic register across text types: The case of modern Hebrew. *Pragmatics and Cognition*, 17 (1), pp. 108-145.

Santini, M. (2007). Characterizing genres of *web pages: Genre hybridism and individualization.* In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. Hawai, pp. 71

Scott, M. (1997). PC analysis of key words. *System,* 25 (2), pp. 233–245.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam/Philadelphia: John Benjamins.