Identifying specific textual units of documents taken from large corpora. Comparing methods.

Francesco Pauli[1], Arjuna Tuzzi[2]

fpauli@stat.unipd.it, arjuna.tuzzi@unipd.it

[1]Dipartimento di Scienze Statistiche, via Battisti 241/243 35123 Padova

[2]Dipartimento di Sociologia, via Cesarotti 10/12 35123 Padova

Abstract

Recent softwares for textual analysis often contain procedures aimed at identifying specific features of documents in large corpora in order to distinguish among them ; especially algorithms based on the hypergeometric probabilistic model. This paper attempts to propose some new directions based on bootstrap techniques. The corpus is composed of documents written by different stakeholder actors during the seven preparatory meetings of the first step of the United Nations World Summit on the Information Society (WSIS - Geneve 2003).

Sommario

Al fine di individuare le peculiarità caratteristiche di documenti contenuti in corpora testuali di grandi dimensioni, i software recenti offrono procedure di calcolo che si basano prevalentemente sul modello probabilistico ipergeometrico. Questo contributo propone nuove strade basate su tecniche di tipo boot-strap. Il corpus in analisi è costituito da documenti prodotti da diversi attori (stakeholders) nell'ambito dei sette incontri preliminari della prima fase del Summit Mondiale delle Nazioni Unite sulla Società dell'Informazione (WSIS - Ginevra 2003).

Keywords : text data mining, specific textual units, bootstrap, hypergeometric model, χ^2 test, WSIS.

1. Introduction

The main goal of data mining is to mine new information from data and finding patterns across datasets. Text data mining concerns the application of data mining techniques to unstructured textual data. The possibilities for data mining from large textual corpora are potentially unlimited because texts contain a rich amount of information, but this information is encoded in a form that is difficult to process by means of statistical methods (Bolasco et al., 2005 ; Sirmakessis, 2004). Text data mining techniques are more useful the larger the corpora, when a direct reading of the texts becomes unfeasible. The identification of textual units (TUs) useful to describe specific features of documents (or groups of documents) in large corpora in order to distinguish among them is one of the most important challenge which arise in this context. Recent softwares for textual analysis often contain procedures for this purpose ; especially algorithms based on the hypergeometric probabilistic model (Lafon, 1980 ; Lebart et al., 1998). Some contributions have already discussed the limits of this approach (Tuzzi et Tweedie, 2000) and proposed supplementary analysis in order to test the strength of conclusions about differences among documents (or groups of documents) obtained by means of the hypergeometric model. This paper aims at rearticulating the limits of these methods and attempts to propose some more convenient alternatives. We used the

Taltac software (Bolasco et al., 2000) for the codifying procedures of the corpus and the R software (R Development Core Team, 2005) for the statistical analysis.

1.1 The corpus : the role of different actors at WSIS

The United Nations World Summit on the Information Society (WSIS), which ended in Tunis last november 2005, has been an important global communicative space, with innovative aspects, both in content and process. The idea of this Summit originated at the International Telecommunication Unit (ITU) Plenipotentiary Conference in Minneapolis in 1998 and the Summit was organized by the ITU in cooperation with other partners : Heads of State, UN agencies, particularly UNESCO and the Information and Communication Technologies (ICT) Task Force, as well as other actors called "stakeholders" (Civil Society and private sector). The preparatory process of the first step of the WSIS process (concluded in Geneve, December 2003) was composed of a series of events : three open-ended preparatory committees (prepcoms), an informal meeting, an intersessional meeting, a number of regional conferences, and so on. The Summit aimed at developing a common vision of the Information Society and at drawing up a strategic plan of action with focus on three main topics : i) visions (meaning the need to develop a common understanding of the information society) ; ii) access (the need to promote the access of all the world's inhabitants to ICTs as well as to skills and knowledge useful to use them) ; iii) applications (in relation to the concrete development goals of the UN Millennium Declaration). The WSIS offered an international stage for two relevant political (communication) debates : the (expected) political negotiation among official delegates for the definition of agreed upon positions to be inserted in the final documents and the (less expected) emergence of a dialogue among different stakeholders. The whole process has therefore created a "world of words" (Padovani et Tuzzi, 2004, 2005a, 2005b).

The WSIS offered a meaningful opportunity to observe the transformations of global communication governance and different actors' impact on global politics. In this paper 41 documents (table 1) which have been elaborated by the different stakeholders involved in the negotiation as contributions to the official process are analyzed in order to reconstruct the WSIS history and to understand the underlying learning process. Documents elaborated by the Civil Society Coordinating Group (CSCG), the Civil Society Content and Theme Group (CSCT) and the Coordinating Committee of Business Interlocutors (CCBI) ; the Official documents ; the regional contributions (EU, Bamako, Bucharest : Beirut ; Tokio and Bavaro) and the Samassekou's paper are clustered according to the seven main preparatory phases of the WSIS : the three official Preparatory committees (Prepcom1 - July 2002, Prepcom2 - February 2003, Prepcom3 September 2003) : the Informal meeting (November 2002) ; the Interesessional meeting (July 2003), the non-official Preparatory committee (Prepcom3A - November 2003) and the Geneve Summit (December 2003). The corpus can be considered large since it is 1.4 MB in plain ASCII.

Num	n. docs	Phase	Authors
1 - 5	5	Prepcom1	Official proposed themes : CCBI contribution ; CSCG contribution : CSCG comment ; EU contribution
6 - 9	4	Informal Meeting	Official outcome ; CCBI contribution : CSCT contribution : EU contribution
10 - 22	13	Prepcom2	Official draft declaration ; Official draft plan of action ; Samassekou's input : CCBI input ; CSCT input plan of action ; CSCT input declaration : CSCG statement ; regional Bamako : regional Bucharest ; regional Beirut ; regional Tokyo ; regional Bavaro ; EU input :
23 - 27	5	Intersessional Meeting	Official draft plan of action ; Official draft declaration ; CCBI input ; CSCT input ; EU input
28 - 33	6	Prepcom3	Official draft declaration ; Official draft plan of action ; CCBI comments plan of action ; CSCT input ; CSCT comments plan of action ; CSCT final document
34 - 38	5	Prepcom3A	Official draft declaration : Official draft plan of action : Official Samassekou's Document : CCBI input ; CSCT statement :
39 - 41	3	Geneve summit	Official declaration : Official plan of action ; CS plenary declaration

***Table 1 :** Description of the corpus (41 documents written by different stakeholders clustered according to seven preparatory phases of the WSIS)*

1.2 Codifying procedure of the corpus

The corpus is a collection of written texts (41 documents) organized according to a grouping criterion (7 phases). The corpus is composed of words which are sequences of letters taken from the alphabet and isolated by means of separators : blanks and punctuationmarks. A word-token (wto) is a particular occurrence of a word-type (wty) in a text. A token instantiates a type (so, for example, the single wty "the" has many tokens in any English text), but there are also many wty that occur only once in a given corpus (hapax legomena). In a first stage of analysis only simple wty could be chosen in order to evaluate the dimensions and the main features of the corpus (Baayen, 2000). However, identifying complex textual units in the vocabulary and codifying the corpus accordingly is sensible (Bolasco, 1999 ; Tuzzi, 2003). Complex textual units are used : i) to increase the amount of information (they carry more information than simple wty) ; ii) to reduce the ambiguity of simple wty (simple wty are more ambiguous because they are totally isolated from their context of usage). In order to improve the corpus we recodified : multi-words ; sequences of words that gain or change meaning if considered as a block and, more generally, sequences that make sense and are repeated several times in the corpus. These operations can be easily performed through the use of Taltac software (Bolasco et al., 2000). Using Taltac procedures we also obtained a list of repeated sequences of words in the corpus. Since most of them were empty (i.e. "and in a", "or the", etc.) redundant (i.e. "cultural and", "cultural and linguistic", "and linguistic diversity", "linguistic diversity" etc.), or incomplete (i.e. "persons with", "countries with economies in") we selected the most informative ones according to the Morrone's statistical IS index (Bolasco et al., 2000), combining this index with a qualitative manual control of the list in order to obtain a new list of the best sequences. The final list was used for the lexicalization procedure. This means that, for example, a repeated sequence such as "countries with economies in transition" was recognized in the corpus as a single textual unit. After the lexicalization procedure simple wty (i.e. "governance"), multi-words (i.e. "civil society"), and repeated sequences (i.e. "marginalized and vulnerable groups") become textual units (TUs) and appear together in the vocabulary. The entire corpus includes a total of *N* =161 465 Tus (corpus dimension in terms of total occurrences and number of statistical textual units).

dimensions	Phase1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6	Phase7
N	19094	5317	35702	31265	38012	15655	16420
V	3506	1463	4711	4801	5063	3160	3709

Table 2 : *Dimensions of sub-corpora (seven phases)*

The list of TUs with each frequency includes a total of V =8 755 TUs (vocabulary dimension in terms of different TUs and number of items considered) and is the vocabulary of the corpus. Dimensions N and V of each phase are shown in table 2.

2. TU distribution across phases

The first issue we considered was deciding whether a TU appears homogeneously across the seven phases, or if it appears mostly in a subset of phases. A TU which is homogeneously distributed across phases should appear in each phase with a frequency roughly proportional to the N-dimension of the phase (as in table 2). In other words we must deal with the issue of testing the difference between the frequency distribution of a certain TU across phases and the expected distribution which implies frequencies (probabilities) proportional to the N-dimension of the phases in terms of total occurrences. The usual (traditional) solution is to employ Fisher (χ^2) test (see Casella et Berger, 2002). In this section we also explore a bootstrap alternative and compare results.

In the bootstrap alternative the TUs in the whole corpus are resampled with replacement according to the following rule : if *x* is the sample of TUs (that is, *x* is the vector of 161 465 TUs in the corpus) and $n._1,\ldots,n._7$ are the N-dimensions of the seven phases, we resample with replacement from *x* to form a vector *x** of the same length : the first $n._1$ elements of *x* are then the bootstrap resample for the first phase, the second $n._2$ are the resample for the second phase and so on. We have then a bootstrap sample of TU frequencies and so, for each iteration, we build a (lexical) contingency table (TUs×phases). As a measure of the specificity of a row, we compute the maximum of the absolute differences between the observed (absolute) frequencies and the expected (absolute) frequencies calculated under the assumption of even distribution of TUs among documents. In other words, if $n._i$ represents the frequency of the TU in the whole corpus and $n._j$ represents the dimension of phase *j*, and *N* (= *n..*) the dimension of the corpus, expected frequencies $\hat{n}_{ij}$ are given by $n_i n._j /N$. We compute the distance between each bootstrap distribution *F** and the expected distribution *F* by

$$d(F^*,\hat{F}) = \max_i \left\{ n_i^* - \hat{n}_i \big| \right\}$$

and compare these with the corresponding distance between the observed distribution $\tilde{F}_{obs}$ and $F : d(\tilde{F}_{obs},\hat{F})$. The bootstrap *p*-value is then given by

$$\frac{1}{B}\sum_{b=1}^{B}\left| d(F^{*b},\hat{F}) > d(\tilde{F}_{obs},\hat{F}) \right|$$

where *B* is the number of bootstrap replications.

We choose resampling with replacement in order to avoid low frequency TUs to appear in all bootstrap samples. The resampling scheme is such that the textual units' frequency in the bootstrap samples changes, that is, the row total in the (lexical) contingency table are not held fixed. This choice is, we believe, particularly appropriate in order to deal with low frequency TUs. If we resample without replacement, a TU which appears once in a corpus of length 10 000 would be bounded to appear in each bootstrap sample and would appear in phase j with a frequency proportional to the length of the group.

We compared the results according to bootstrap with replacement and bootstrap without replacement and noted that results are fairly similar for TUs with not too low frequency (see figure 1), it emerges that tests based on resampling with replacement on average leads to lower significance levels, which means that we identify more specific TUs.

2.1. Results

We compare bootstrap p-values and those based on χ^2 test for TUs with frequency 2 in figure 2. It is worth distinguishing some groups of *p*-values in order to compare the two methods. Four groups are depicted in figure 2 :

a) TUs for which bootstrap *p*-value is less than 0.2 occur twice in the same phase, χ^2 *p*-value takes 7 possible values depending on which phase the TU belongs to, due to different length of the 7 phases.

b) TUs for which bootstrap *p*-value belongs to [0.2, 0.3] occur once in the shortest phase (phase 2 - Informal meeting, 5 317 TUs) and once in any of the other phases.

c) TUs for which bootstrap *p*-value belongs to [0.4, 0.6] and χ^2 *p*-value is less than 0.25 do not occur in the shortest phase nor in the three longest ones.

c+d) TUs for which bootstrap *p*-value belongs to [0.4, 0.6] do not occur in the shortest phase.

The other observed pairs are not of interest since they are far from any reasonable significance level.

All TUs occurring once in the shortest phase are not significant according to bootstrap test, while they are significant according to χ^2 depending on which phase the second occurrence belongs to. When a TU occurs twice in the same phase, the χ^2 *p*-value varies according to the length of the phase, the bootstrap *p*-values also varies (clearly), but to a lesser extent. In other words, χ^2 tests results are heavily driven by the dimension of the texts. In table 3 an example of numerical results is shown.

3. The importance of understanding issues emerged in the seven phases

Writing documents during theWSIS preparatory process, governments, private companies and particularly CS organizations had an opportunity to test their potential impact in a global setting. It is important to underline that CS praxis in the trans-national environment presented meaningful variations : a plurality of manifestations of formal and informal character, institutionalized relations as well as spontaneous self-organization, habits of dialogue with formal institutions together with strong expressions of contentious politics, etc.

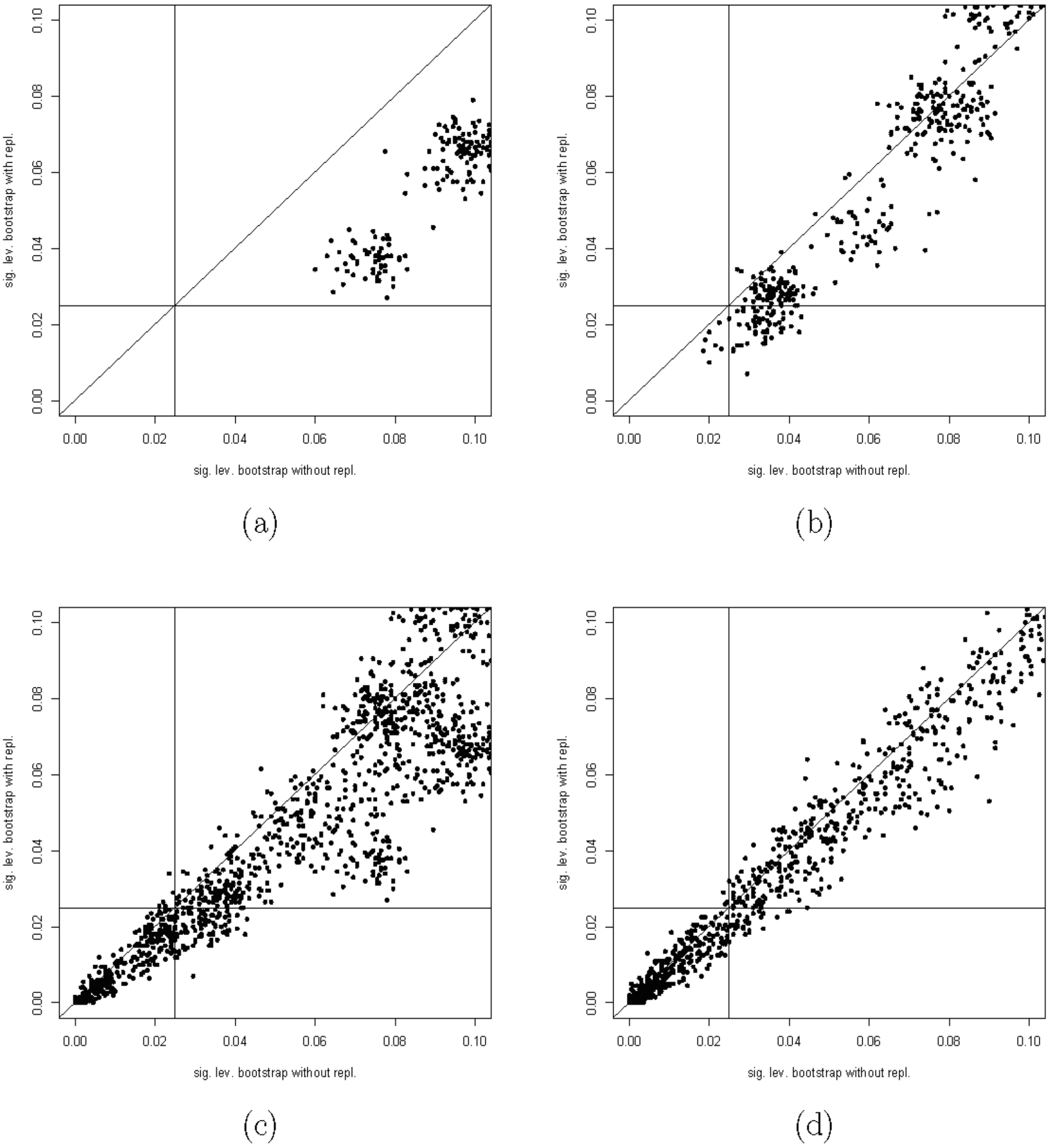
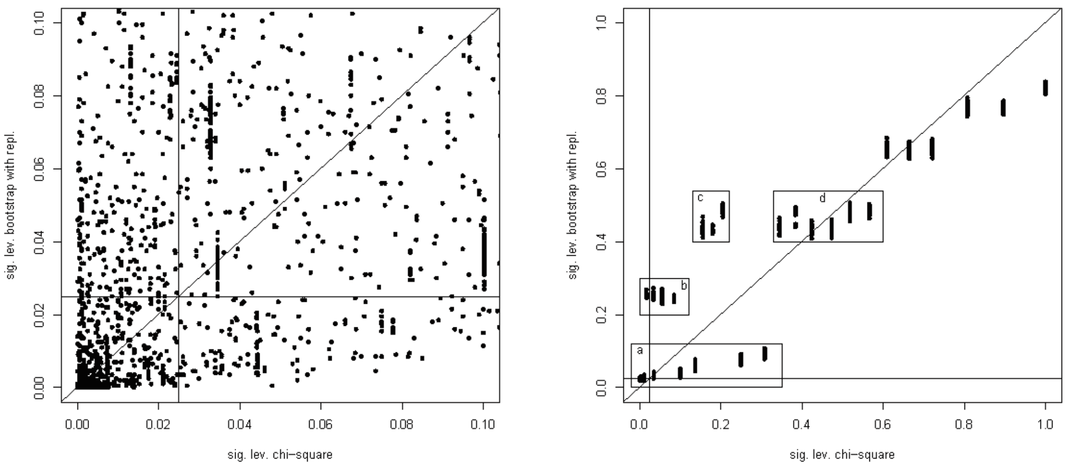


***Figure 1 :** Comparison between bootstrap p-values obtained by resampling with replacement and without replacement for textual units with frequency 1 (a), 2 (b), less than 7 (c), more than 7 (d).*



(a) (b)

Figure 2 : *Comparison between bootstrap p-values obtained resampling with replacement and χ^2 p-values for textual units of any frequency (a) with frequency 2 (b).*

The 41 documents deriving from this multi-stakeholder dialogue are analyzed in order to reconstruct the WSIS history across the seven phases and to have a deeper understanding of the learning process. In order to describe the seven phases it is important : 1) to identify specific TUs which distinguish among phases ; 2) to understand if a specific TU is representative of a phase as a whole rather than specific to a small subset of documents of that phase.

3.1. The hypergeometric model and its limits

In order to describe phases it is possible to use the traditional “characteristic” TUs method (Lebart et al., 1998) based on the hypergeometric model (Lafon, 1980). By means of a probability of over-usage it can detect which elements are used frequently inside a phase (as well as which elements are used rarely) and all TUs which show a high probability of over-usage for a phase can be considered “specific” to that phase with reference to the others.

The occurence of a TU in a phase is not a simple attribute because a phase is composed of different documents and the occurence of a TU in the phase is the sum of the occurrences for each document assigned to that phase. If a TU occurs a great deal more in a small subset of documents than in the rest of the other documents assigned to the same phase and in the rest of the corpus, it is erroneously considered “specific” for the entire phase and not only for that subset of documents. A measure of the dispersion (Baayen, 1996) of the TU inside the phase is important to test if a TU is restricted to a small subset of documents or is well spread out over all the documents of the phase. Only in the second case a TU detected as “specific” by means of the hypergeometric model can be actually considered “specific” for the phase.

TU	Corpus	*p*-value		Frequencies in the seven phases						
	freq	χ^2	bts	1	2	3	4	5	6	7
the	11402	0.0005	0.0000	1423	463	2682	2101	2490	1145	1098
and	10041	0.0005	0.0000	922	286	2260	2010	2349	1000	1214
of	7546	0.0005	0.0000	842	276	1790	1415	1630	750	843
in	4201	0.0005	0.0090	456	113	966	735	990	460	481
to	2827	0.0070	0.0030	380	83	631	579	594	262	298
a	2564	0.0010	0.0310	356	80	523	523	633	232	217
for	2515	0.0350	0.0390	301	99	606	459	591	213	246
that	1495	0.0025	0.0015	195	56	274	272	381	176	141
on	1214	0.0005	0.0150	177	53	241	193	306	123	121
as	1139	0.0665	0.1500	149	40	277	193	243	108	129
with	1027	0.0625	0.2110	127	17	218	223	241	97	104
by	969	0.0830	0.2540	137	41	210	174	220	79	108
is	967	0.0005	0.0000	163	32	222	158	243	80	69
an	823	0.0070	0.0455	108	37	169	131	194	102	82
are	778	0.0005	0.0080	109	42	148	115	177	92	95
development	742	0.1924	0.1035	73	26	188	147	179	61	68
information society	708	0.0005	0.0060	52	50	168	123	149	103	63
all	651	0.0210	0.0200	48	22	163	120	156	67	75
this	645	0.0005	0.0000	70	26	131	83	208	62	65
their	636	0.0005	0.0020	45	7	126	129	156	74	99
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
assessed	7	0.2041	0.0675	0	0	1	4	2	0	0
evaluated	7	0.6101	0.6240	0	0	2	1	2	0	2
industry-led	7	0.0313	0.0255	4	0	0	1	2	0	0
broadly	7	0.3434	0.3975	2	0	0	3	1	1	0
brought	7	0.0913	0.0750	0	0	3	4	0	0	0
interfaces	7	0.3914	0.4655	2	0	0	0	3	1	1
ethnic	7	0.6101	0.6255	0	0	2	1	2	0	2
marginalized urban areas	7	0.4500	0.4415	0	0	0	2	2	2	1
eliminating	7	0.1385	0.1745	3	0	2	2	0	0	0
marginalized and vulnerable groups	7	0.0464	0.1315	0	0	0	1	1	3	2
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
probably	2	0.4728	0.4560	0	0	0	0	1	1	0
viii	2	0.3073	0.0980	0	0	0	0	2	0	0
tensions	2	0.5177	0.4640	0	0	1	0	0	0	1
to form	2	0.1553	0.4430	0	0	0	0	0	1	1
textbooks	2	0.2495	0.0760	0	0	2	0	0	0	0
representativeness	2	0.1364	0.0535	0	0	0	2	0	0	0
test	2	0.6086	0.6580	1	0	0	1	0	0	0
phase of the wsis	2	0.0345	0.0345	0	0	0	0	0	0	2
arrive	1	0.7633	0.3550	0	0	1	0	0	0	0
located	1	0.1312	0.0940	0	0	0	0	0	1	0
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

***Table 3 :** χ^2 and bootstrap p-values for a selection of TUs.*

Results obtained by means of resampling-based tests aimed at identifying TUs which are significantly under-dispersed inside a group of documents developed by Baayen (1996) and Tuzzi et Tweedie (2000) show that the hypergeometric model alone is not sufficient to draw general conclusions about differences among groups of documents. However, these tests, based on Monte Carlo algorithms, are time consuming to the extent that when corpora are large they may be unfeasible and, furthermore, in general they are not robust for low frequency TUs.

In the section below we propose an alternative based again on homogeneity bootstrap tests.

3.2. Specific of a phase or specific of a subset of documents of the phase ?

From a technical point of view this issue is no different than that we dealt with in section 2 where the phase we are investigating in depth plays the role of the corpus and its documents play the role of the phases. As in section 2, we consider as alternative techniques the χ^2 test and the bootstrap with replacement, where the latter is implemented as explained in section 2.

3.3. Results

As far as phase 4 is concerned, for example, the TU “should be promoted” which is significant according to hypergeometric test (p-value equal to 0.001) is not representative of the whole phase 4, since it is over represented in some documents. On the contrary, the TU “indicators”, which is significant according to hypergeometric test (*p*-value equal to 0.013) is representative of the whole phase 4, since it is evenly represented in all documents. Only this second TU is actually useful to summarize the specific features of phase 4.

TU	corpus	Ph.4	cell	homog		sub-phase												
	freq	freq	hyp	χ^2	bts	1	2	3	4	5	6	7	8	9	10	11	12	13
through	427	101	.016	.003	.009	9	15	8	4	20	1	6	4	7	14	12	11	8
creation	204	50	.041	.030	.059	5	6	4	2	14	2	5	2	3	4	1	4	5
local	186	48	.019	.002	.005	6	1	2	1	13	0	5	12	1	3	1	8	6
better	77	23	.018	.013	.025	3	2	4	0	2	0	2	2	2	2	3	0	4
local content	65	22	.004	.038	.070	3	2	3	0	4	0	0	1	2	0	2	0	4
e-business	57	17	.038	.004	.012	2	1	2	0	0	0	0	0	2	8	2	0	5
indicators	47	16	.013	.041	.067	2	2	0	0	2	0	0	0	1	5	0	1	3
should be promoted	46	19	.001	.000	.001	3	5	1	0	4	1	0	1	0	0	4	1	1
help	45	15	.019	.005	.011	0	3	0	1	6	0	0	1	1	1	2	2	1
entrepreneurship	28	10	.032	.030	.050	0	1	1	3	1	0	1	0	1	0	0	1	2
personal	26	9	.050	.006	.011	0	2	1	0	4	1	0	1	0	0	2	0	1
reports	10	5	.029	.049	.066	0	0	0	0	3	0	0	1	0	0	0	0	1

Table 4 : *Test of the distribution of TU internal to the phase for a selection of TU in phase 4 : Intersessional meeting.*

In order to show the power of the proposed method we list below a selection of TUs that are phase significant meaning that are both specific for the phase and evenly distributed among the document within the phase. These TUs are part of a discourse which emerge as typical of that phase, this means of that historical period, and represent issues on which a substantial agreement exists among documents, that is, among different stakeholder, or controversial issues that are of great relevance in all documents.

- Phase 1, Prepcom1 : advertising, ahead, approach, border, consolidation, consumer trust, contents, corresponds, customer, discussion, Doha, economic, edition, fraud, global electronic commerce, globally, infrastructures, interest, internal, internationally, interoperability, jurisdiction, jurisdiction and applicable law, maternal, nature, networks, new, non-profit, open-ended, signature, suppliers, to continue, treatment, unequally, urgently, 2001.

• Phase 2, Informal Meeting : chairman, divides, following, human rights, info, informal, information network security, international fora, issue, negotiation, network security, paper, preliminary, Prepcom, proposal, proposed, sciences, seize, spectrum, sub-committee, supporting, universal, wide, WSIS, yesterday.

• Phase 3, Prepcom2 : access to information and communication technologies, aimed, constraints, countries, democracy, e-learning, enterprises, existence, formation, funding, healthcare, markets, private sector and civil society, regional and subregional, regional conference, regulatory and policy frameworks, sector, setting up, shaping, steps, ubiquitous access to information.

• Phase 4, Intersessional Meeting : e-health, interconnection, must, network, opensource software, should be used, web.

• Phase 5, Prepcom3 : amendments, by 2005, by 2010, by 2015, create, disability, ensure, environment, establish systems, ethnicity, expense, general, mention, natural disasters, race, section, should, women and girls, would be.

• Phase 6, Prepcom3A : bottom-up, building the information society, circuit, dignity, draft, inclusive information society, key principles, operator, other stakeholders, plan of action, satisfied, someone, special needs, sustainable development, these technologies, this declaration, time, well-being.

• Phase 7, Geneve summit : citizenry, collective, communication societies, conflict situations, disadvantaged and vulnerable groups, efforts, freely, human knowledge, information and communications, instead, knowledge societies, participatory, pluralistic, public domain, this implies, to encourage innovation and creativity, we recognise.

If few or none TUs are phase significant, we could assume that there are no features distinctive of a phase and at the same time common to all stakeholders within that phase.

4. Conclusions

From a "contextual" point of view, looking at content and political implications, this analysis offers a new and wider perspective to better understand the chronological development of the WSIS discourse up to the Geneva Summit and shows how the agenda has been trasformed and enlarged over time. At the very beginning of the process (Prepcom1) issues related to technological and economic/commercial prevail. In the following phase (Informal Meeting) different issues emerged as central : reference to the digital divide and to network security alongside with an interest for human rights. During the third phase (Prepcom2) the problem of "divides" seems to be translated into more positive terms as documents pay more attention on how to solve the problem (recurrent reference to "access"). In this phase actors (stakeholders) and different political levels are also explicitly mentioned. In the documents from the fourth and fifth phases (Intersessional Meeting and Prepcom3 respectively) the discourse is centred on ICT applications while issues are addressed in a more specific manner (e.g. gender, enviroment, disabilities, etc.). The sixth phase (Prepcom3A) is interesting since the meeting was organized to allow for further negotiation because there was no agreement on the final draft documents yet during this phase it seems that there is a shared recognition of the relevance of the value dimension and this finding suggests that further investigation would be appropriate. In the last phase (Geneve summit) a more articulated vision of the information society is shared. Findings from former analyses demostrate several differences among documents (Padovani et Tuzzi, 2004) but it is interesting to notice that a more complex articulation of the discourse is the result of the learning process in which all speakers have been involved.

From a “methodological” point of view, as already shown in Tuzzi et Tweedie (2000) and Baayen (1996) testing for specificity of a TU across phases is not enough to draw general conclusions about differences among phases because these analysis need to be deepened in order to ensure that the considered TUs are representative of the phases as a whole and their specificity is not due to a small subset of documents. We suggest that in order to draw sensible conclusions, these should be based on a table similar to table 4, including results of hypergeometric tests, χ^2 and bootstrap tests for homogeneity within each phase.

The test procedure we proposed provides the same answers as the above ones, but it is less time consuming and it seems not suffer of lack of robustness for low frequency TUs.

Acknowledgements

We wish to thank Claudia Padovani (Department of Historical and Political Studies of Padua) who supplied us with these documents collected for her scientific works on the WSIS and for her useful insights.

References

Baayen H.R. (2000). *Word Frequency Distributions. Exploring Quantitative Aspects of Lexical Structure*. Kluwer Academic Pub.

Bolasco S. (1999). *Analisi multidimensionale dei dati*. Carocci, Roma.

Bolasco S., Canzonetti A., Capo F. M. (2005, eds). *Text Mining. Uno strumento strategico per imprese e istituzioni*. CISU, Roma

Bolasco S., Baiocchi F., Morrone A. (2000). *TALTAC Versione 1.0*. CISU, Roma.

Casella G., Berger R.L. (2002). *Statistical inference*. Duxbury, Pacific Grove, USA.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, I : 127-65.

Lebart L., Salem A., Berry L. (1998). *Exploring Textual Data*. Kluwer-Academic Pub.

Padovani C., Tuzzi A. (2004). The WSIS as a World of Words : Building a Common Vision of the Information Society ? Continuum, *Journal of Media & Cultural Studies,* Vol. 18, No. 3, September 2004 : 360-379.

Padovani C., Tuzzi A. (2005a). Communication Governance And The Role Of Civil Society. Reflections on Participation and the Changing Scope of Political Action. In Servaes J. and Carpentier N. (eds), *Deconstructing WSIS, Towards a Sustainable Agenda for the Future Information Society*, Intellect Books, Bristol.

Padovani C., Tuzzi A. (2005b). La comunicazione politica internazionale e il ruolo della società civile : reti comunicative nel Summit Mondiale sulla Societ`a dell’Informazione. Riflessioni su governance, partecipazione e trasformazioni della politica a partire dall’analisi, lessico-testuale dei documenti. Redes.com, vol. 2 : 307-35.

R Development Core Team (2005). R : A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria, http://www.R-project.org.

Sirmakessis S. (2004). Text Mining and its Applications. *Results of the NEMIS Launch Conference,* Springer, Berlin-Tokio.

Tuzzi A., Tweedie F.J. (2000). The Best of Both Worlds : Combining MOCAR and MCDISP. In Rajman M. e Chappelier J.C., *JADT 2000 5es Journées internationales d'Analyse statistique des Données Textuelles*, EPFL ed., Vol. 1 : 271-76.

Tuzzi A. (2003). *L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*. Carocci, Roma.