

-
-



Modeling Noun-Phrases Dynamics in Specialized Text Collections

Séminaire IFRIS

Vendredi 16 Octobre, 2009, IFRIS-Marne la Vallée

Nicolas Turenne

INRA – Jouy-en-Josas centre

Research Context

- INRA
 - Main topics: **Fundamental Biology, Food, Environment, and New Techniques for Agriculture**
 - National Research Organization : **20 research centres, 400 research units, 9000 members** (3000 full time staff research scientists in various fields not only basic biology, also in ‘Sciences for Modeling’, Research Policies Economy and Applied Biology).
- My research interest centre
 - **Knowledge Engineering from Texts** and applications to *Sociology of Science and Bioinformatics*
 - **Document Content** plays a key role as a knowledge indicator
 - **Evolution of Content**, i.e. time, plays a key role as a structural parameter

Plan of the presentation

PART 1

- State of the art
lexical change analysis, sequential analysis, linguistic distributions

PART 2

- Content-words: experiments
- Content-words: modeling

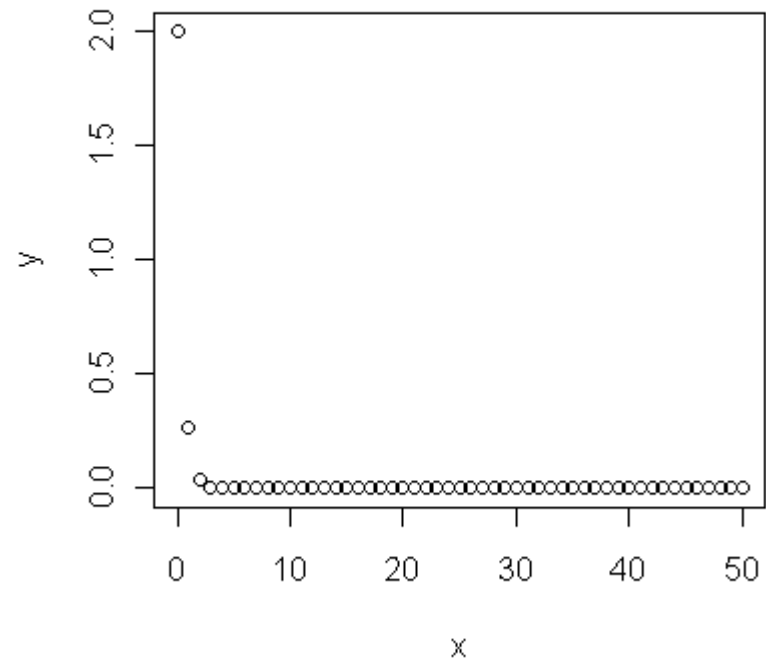
PART 3

- Dependencies: experiments
- Dependencies: modeling

Word Distribution - Zipf

$$y = \frac{n}{x}$$

- Where y = intensity, x = frequency, n = proportionality factor.
- G.K. Zipf published his doctoral dissertation, which dealt with changes of sounds phonemes.



Word evolution over time

- Lexical Change Analysis
- Sequential Analysis

Lexical Change Analysis

- Successive editions of dictionaries

[Dubois, 1962;
Neuhaus, 1973]

- **Corpora within different periods**

[Kroeber and Chrétien, 1937;
Swadesh 1952, 1955;
Labov, 1980 ;
Embleton, 1986;
Baayen and Renouf , 1996]

Lexical Change Analysis

- Glottochronology (historical divergence of languages)

- Radioactivity decay Model for words – Swadesh approach

$$N(t) = N(0).e^{-\eta t}$$

- Tree reconstruction which is based on techniques used in biology – Embleton approach

Lexical Change Analysis

- Arapov and Cherc Approach

$$n_i(t,0) = 1 - e^{-\eta_i v_i}$$

$$F_i(t,0) = n e^{-\eta_i t}$$

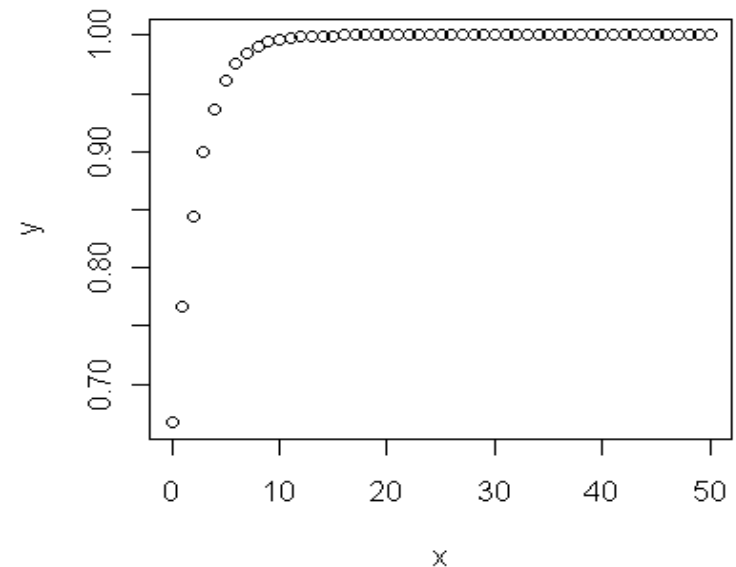
$$v_i(t_1, t_2) = e^{-\eta_i t_1} - e^{-\eta_i t_2}$$

- probability of a word to be handed to a follower-up lexicon
- Probability increases with age and decreases with the square root of frequency
- F_i is the number of words which prevail in the rank class i after time t , n the size of the rank class, and η the (variable) decay rate
- distribution of the loans from a source in the time interval $[t_1, t_2]$

Lexical Change Analysis

- Piotrowski Law
 - development (increase and/or decrease) of the portion of new units or forms over time. This law is a typical growth process and can be derived from a simple differential equation with the solution
 - where $a, b > 0$
- Polikarpov (1993) developed a word life cycle, for building a theory of the organisation and historical development of language systems as a whole.

$$y = \frac{1}{1 + a \cdot e^{-b \cdot x}}$$



Sequential Analysis

- Saussure (1916) introduced the notion of syntagmatic axis, analogous to time axis and reflecting text sequence.
- theory of Markov chains (Baum & Petrie, 1966).
- information theory calculating possible arrangement in the line (Shannon, 1948)
- time-series analysis (Pawlowski, 1997; Pawlowski, 1999) based on spectral and on ARIMA method (Autoregressive Integrated Moving Average) (Box&Jenkins, 1970; Nurius, 1983)

Sequential Analysis

$$x_t \left(1 - \sum_{i=1}^p a_i B^i\right) = e_t \left(1 - \sum_{k=1}^q b_k B^k\right)$$

- B is the lag operator, the a_i are the parameters of the autoregressive part of the model (AR), the b_k are the parameters moving average part and the e_t are error terms.

$$y_k = E[(X_t - \mu_x)(X_{t+k} - \mu_x)]$$

- y_k is the autocovariance of the series

$$c_k = \frac{1}{N-k} \sum_{t=1}^{N-k} [(X_t - m_x)(X_{t+k} - m_x)]$$

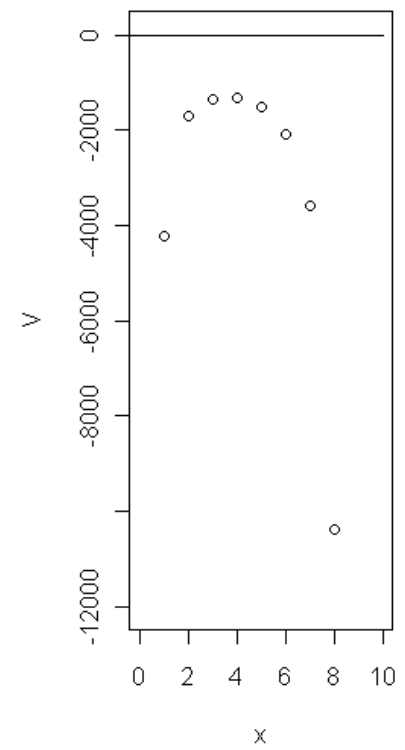
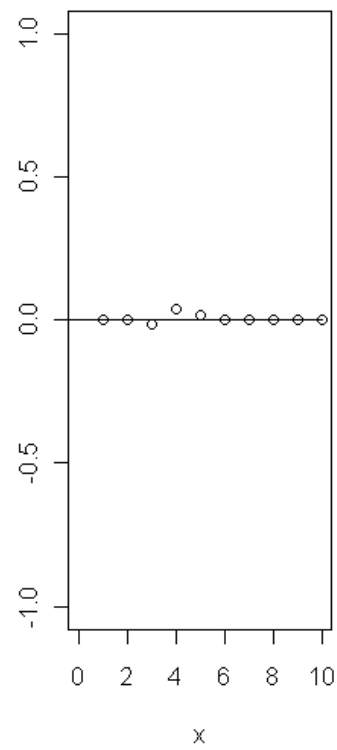
- estimated by c_k

Sequential Analysis

- Autocorelation of the series
- Estimated by
- Pawlowski's law predicts autocorrelation of phoneme when a text segment is interpreted as a time series of units with a seasonal lag

$$r_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma_x^2}$$

$$\bar{r}_k = \frac{C_k}{S_x^2}$$



$$x_t \left(1 - \sum_{i=1}^p a_i B^i\right) \left(1 - \sum_{j=1}^p a_{sj} B^{sj}\right) = e_t \left(1 - \sum_{k=1}^q b_k B^k\right) \left(1 - \sum_{l=1}^Q b_{sl} B^{sl}\right)$$

- The law is expressed by the explained variance

$$V_e = 100\% \left(1 - \frac{S_r^2}{S_{obs}^2}\right)$$

Linguistic Distributions

- Lots of linguistic laws
 - more than 40 (not only Zipf)
- Distribution modeling closed to our content-words modeling
 - linear shape laws
- Distribution modeling closed to our dependencies modeling
 - Complexity in Synergetics Theory

Linear-Shape Laws

- Oono's law (Oono, 1956) states that the proportion between the numbers of all parts of speech keeps the same over time in the lexicon of a language, although, typically, the lexicon size grows .

- Hrebicek (2005) studied distribution of co-references. He supposes that :

$$dz = a.k.dw$$
 where w is a text cohesion parameter it should be defined as the relative mean frequency of words ($w=v/n$). He postulates also that

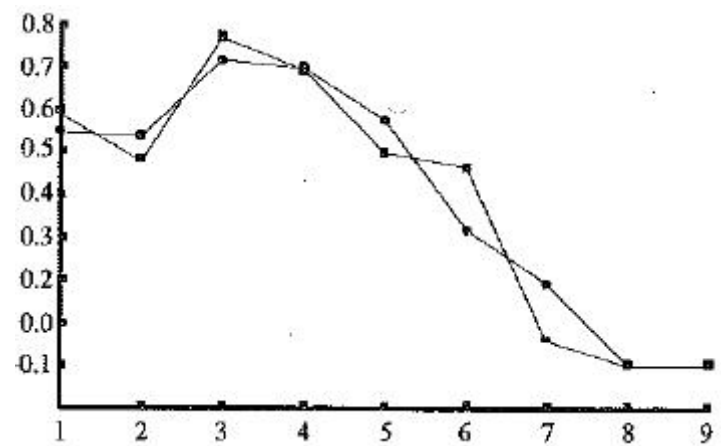
$$dz = a.w.dk$$

Complexity

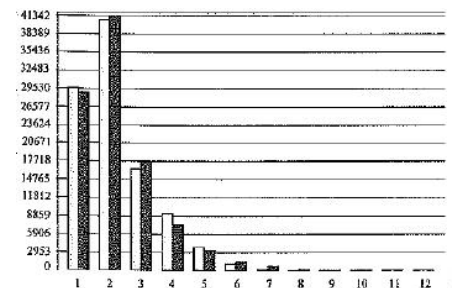
- Kohler (1999) [synergetic linguistics theory] observed that component dependency upon their position do not appear to be shaped as classical power law such as between other parameter (frequency, length, lexicon size, minimization of production effort, minimization of decoding effort...)

$$P_x = \frac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}} q^x P_0$$

where $P_0^{-1} = {}_2F^1(k, 1; m; q)$ the hypergeometric function as nor min g cons tan t.



- fitting the Hyper-Pascal distribution to the complexity (Suzanne corpus)



Evolution of “Content-Words” in Corpora of Biology

- Corpora

- **Pubmed** <http://www.ncbi.nlm.nih.gov/pubmed/> 20 millions of documents

- **CorpusM**

- 34,529 documents

- topic: Mouse species and its embryo development

- 276,100 sentences

- dispatched into 7 time intervals [1963-1968, 1968-1987, 1987-1994, 1994-1999, 1999-2002, 2002-2005, 2005-2007].

- **CorpusH**

- 77,333 documents

- topic: Human species, embryo, placenta and cancer

- 515,500 sentences

- dispatched into 12 time intervals [1963-1970, 1970-1979, 1979-1985, 1985-1990, 1990-1993, 1993-1996, 1996-1998, 1998-2000, 2000-2002, 2002-2004, 2004-2005, 2005-2007].

Evolution of “Content-Words” in Corpora of Biology

- **Named Entities Extraction**

- Extractors

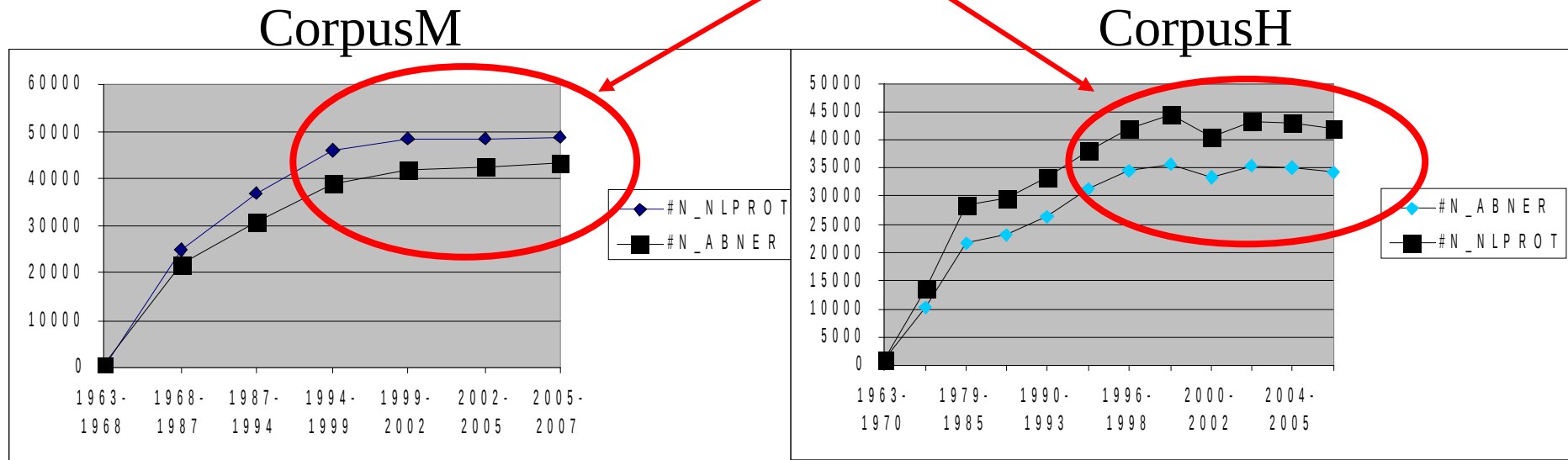
- Abner (Settles, 2005)
- Nlprot (Mika and Rost, 2004)
- machine learning with a training corpus:
- Nlprot: biology name dictionaries and support vector machine classifiers
- Abner : conditional random fields models

- Noun-phrases Extraction

- CorpusM 60,611 NPs (Abner) & 42,427 (Nlprot)
- CorpusH 82,903 NPs (Abner) & 48,086 (Nlprot)

Experimental Results , first observation

plot : number of gene/protein names found by time points => **we can see a plateau**

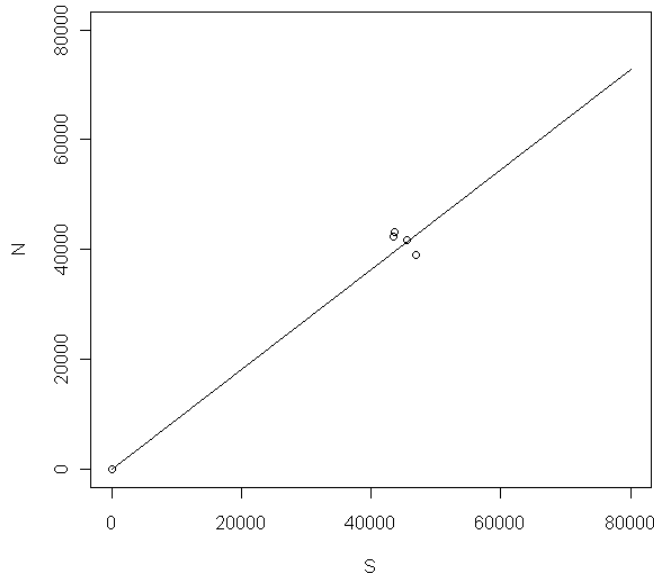


Extractor influence

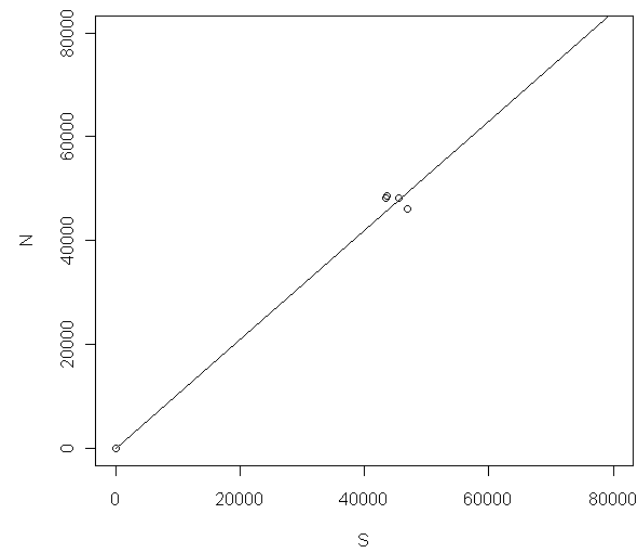
- we observe a linearity between the number of names entities and number of sentences.

CorpusM

nlprot extractor



abner extractor



Modeling for corpusM

$$N_{\text{CorpusM}}^{\text{abner}} = 0.91 * S$$

Where $N_{\text{CorpusM}}^{\text{abner}}$ is number of named entities.

$$N_{\text{CorpusM}}^{\text{nlprot}} = 1.05 * S$$

Where $N_{\text{CorpusM}}^{\text{nlprot}}$ is number of named entities.

$$\hat{N}_{\text{CorpusM}} = k_{\text{CorpusM}} * S$$

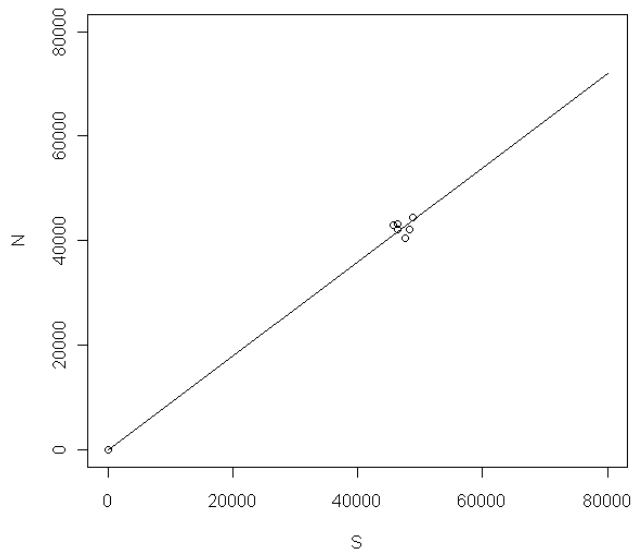
Where \hat{N}_{CorpusM} is mean number of named entities for CorpusM,

$$k_{\text{CorpusM}} = 0.98 \pm 0.07.$$

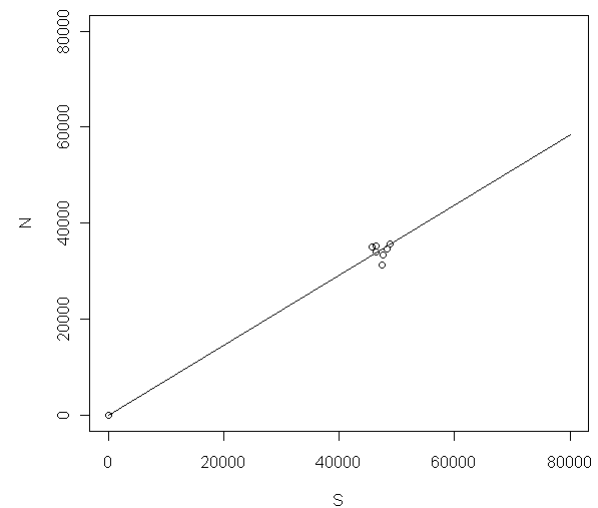
Corpus influence

- same linearity for corpusH.

nlprot extractor



abner extractor



Modeling for CorpusH

$$N_{\text{CorpusH}}^{\text{abner}} = 0.73 * S$$

Where $N_{\text{CorpusH}}^{\text{abner}}$ is number of named entities.

$$N_{\text{CorpusH}}^{\text{nlprot}} = 0.90 * S$$

Where $N_{\text{CorpusH}}^{\text{nlprot}}$ is number of named entities.

$$\hat{N}_{\text{CorpusH}} = k_{\text{CorpusH}} * S$$

Where \hat{N}_{CorpusH} is mean number of named entities for CorpusH,
and $k_{\text{CorpusH}} = 0.81 \pm 0.08$.

Model of “Content-word” Dependencies over time

- **Corpus**

- Pubmed & SCI (web of science)

- title+abstracts+keywords

- CorpusP

- corpus of biology based on a protein called prion (or prp) and responsible of mad-cow disease.
- 6,658 documents.
- set of time intervals, for author names 9 intervals [1985-1986, 1987-1988, 1989-1990, 1991-1992, 1993-1994, 1995-1996, 1997-1998, 1999-2000, 2001-2002], for nouns phrases 7 intervals [1985, 1986, 1987-1988, 1989-1994, 1995, 1996-1997, 1998-2002]

- CorpusE

- corpus of biology is widely focused on plant epidemiology.
- 5,545 documents.
- a step of one year from 1995 to 2005

- CorpusC

- corpus of recent clustering techniques area.
- 982 documents.
- 5 intervals were defined [1991-1994; 1995-1997; 1998-1999, 2000-2001, 2002-2003]

Model of "Content-word" Dependencies over time

- AU North, L. H.
- Wuest, P. J.
- TI The **infection process** and symptom expression of Verticillium disease of Agaricus bisporus.
- AB Interactions between *V. fungicola* and its host *A. bisporus* leading to differing symptom types, including dry bubble, necrotic lesions and stipe blowout, were investigated by observation of naturally infected mushrooms by SEM. There was no evidence of specialized penetration structures or of direct penetration. However, *V. fungicola* was closely associated with the surface of the sporophore and with the internal hyphae of its host. Primordia, developing sporophores and maturing sporophores of 4 cultivars of *A. bisporus* were inoculated with 5 micro l droplets of spore suspensions of *V. fungicola* containing 2 x 10⁵, 2 x 10⁶ or 10 x 10⁶ spores/ml. The type of symptom produced depended on the age of the sporophore at the time of inoculation, inoculum density, and on post-inoculation incubation time. Dry bubble developed after inoculation of primordia, necrotic lesions occurred in all treatments, while stipe blowout developed only in sporophores inoculated with the highest spore concn. The **incidence of sporophores** with symptoms before harvest increased with spore concn. At the lowest spore concn, 90% of the inoculated mushrooms developed symptoms after harvest. It is suggested that the occurrence of infected mushrooms without symptoms at harvest may affect the epidemiology and control of the disease and may lead to reduced eye-appeal of mushrooms at retail markets.
- DE mushrooms; infection; symptoms; plant pathology; plant pathogenic fungi; edible fungi

a document

dependency

a content-word

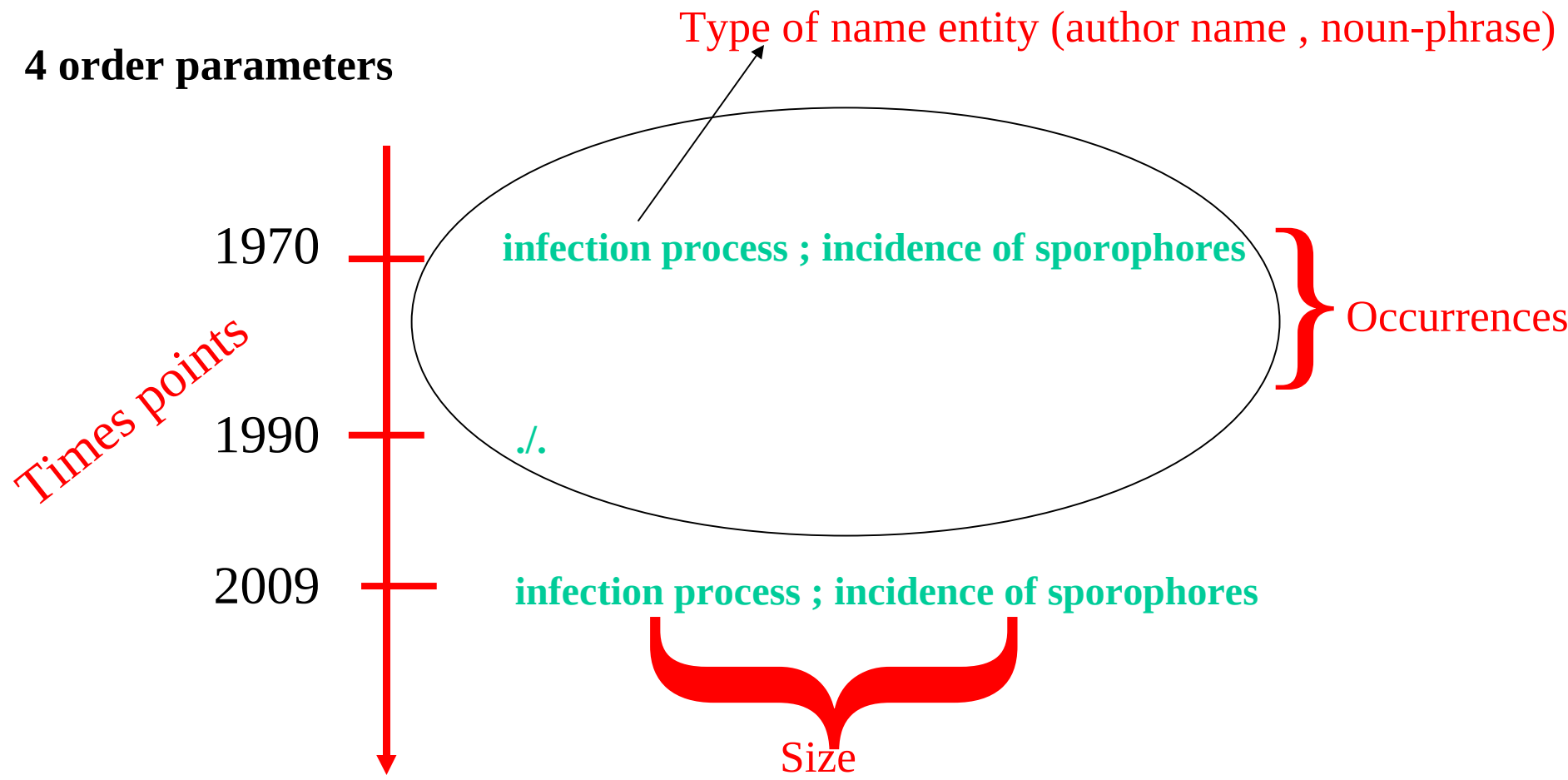
Model of “Content-word” Dependencies over time

“Content-Words” Dependencies Extraction

- **Beluga** (Turenne & Barbier, 2004) to extract
- **Clustering algorithm**
 - sequential pattern extraction algorithm (Agrawal and Srikant, 1994)
 - catches co-occurrence of n words s times:
 - s is called support,
 - n words is called a sequential pattern (i.e. a word cluster),
 - a context is a document.
- **Support** for both has been settled to 3 for CorpusP, and 2 for CorpusC and CorpusE. Such time scale was defined to get an equi distribution for documents. For instance about CorpusC and CorpusE each interval at beginning contains in average 200 documents per interval.
- s is ideally settled to 2 but can be higher to solve computational limit of memory storage.
- Time point $T=1$ means the most recent interval, $T=2$ the immediate successor, etc...

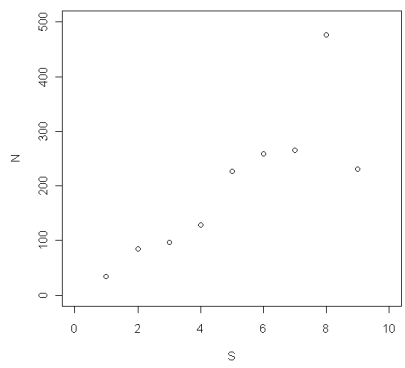
Model of "Content-word" Dependencies over time

4 order parameters

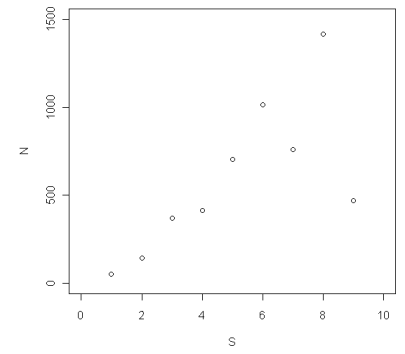


Experimental Results with Authors Names

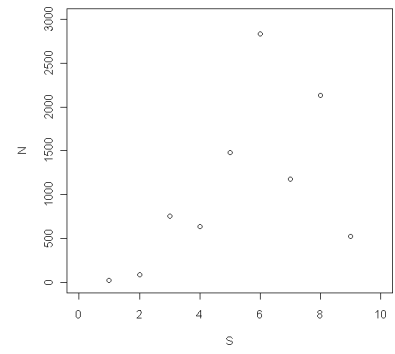
Distribution of patterns over time given a pattern size N for CorpusP
 (a) $N=1$ (b) $N=2$ (c) $N=3$ (d) $N=4$.



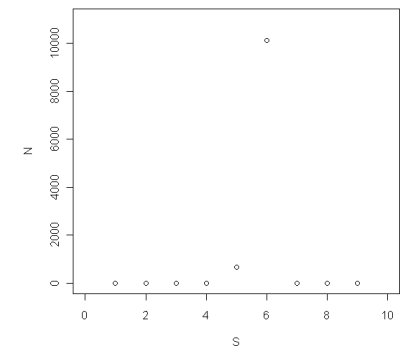
(a)



(b)



(c)

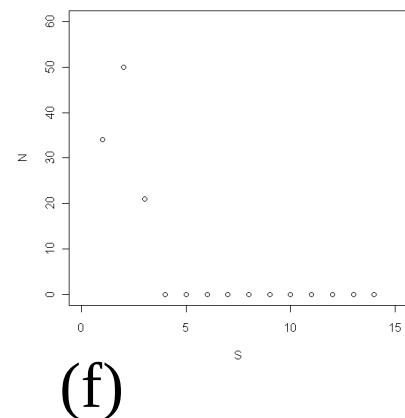
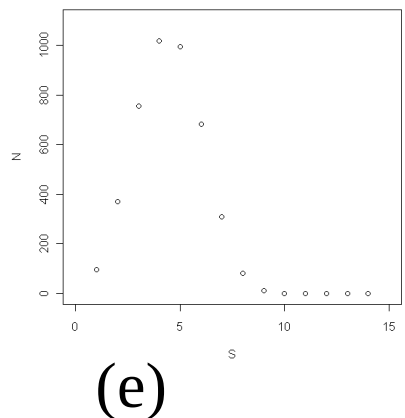
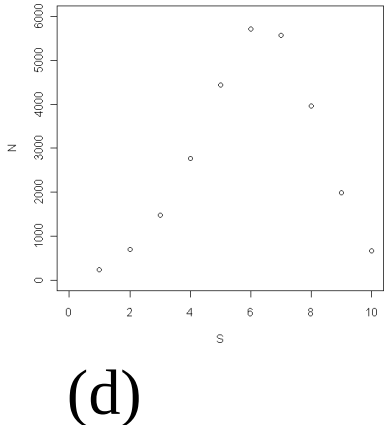
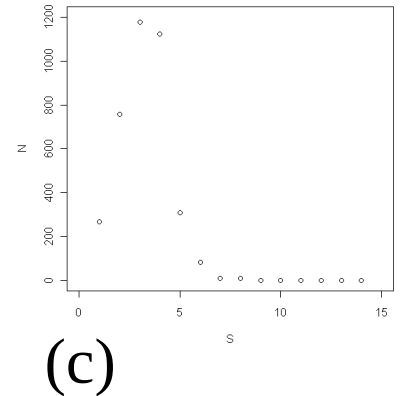
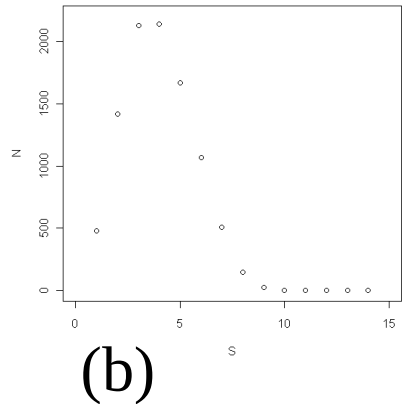
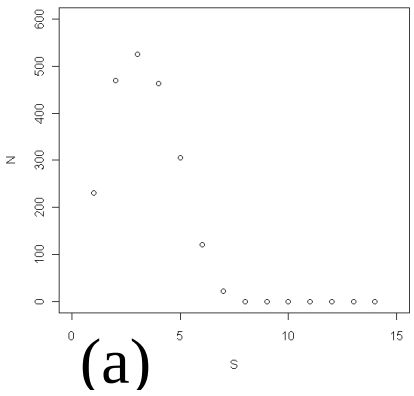


(d)

Experimental Results with Authors Names

Distribution of patterns size given a period T for CorpusP

(a) T=1 (b) T=2 (c) T=3 (d) T=5 (e) T=7 (f) T=9

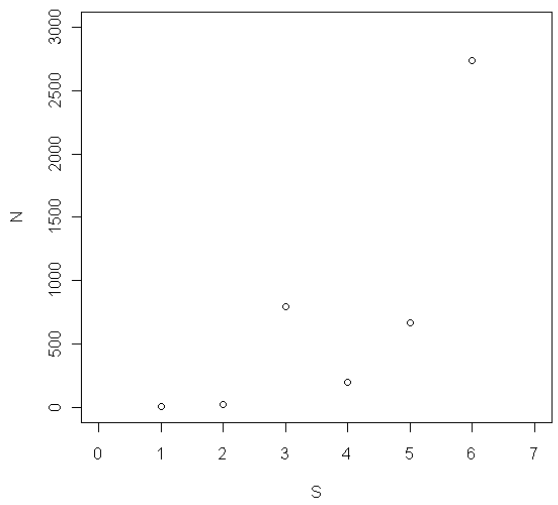


Experimental Results with Authors Names

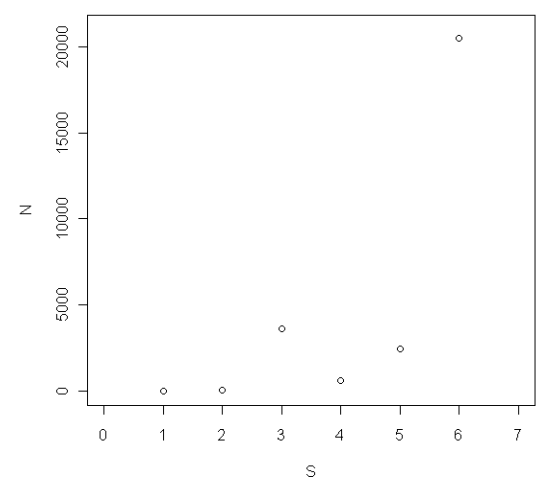
- The profile is very close to the law describes and modeled by (Köhler, 2000) Hyperpascal distribution.
- shift of the distribution occurring at period $T=5$, and leading to an average size of patterns around 5 components, though usually the average was around 3 components.
- hyper-Pascal distribution seems time-dependant but localization of this dependency seems to be unpredictable and domain-dependant

Experimental Results and Modeling with "Content-Words"

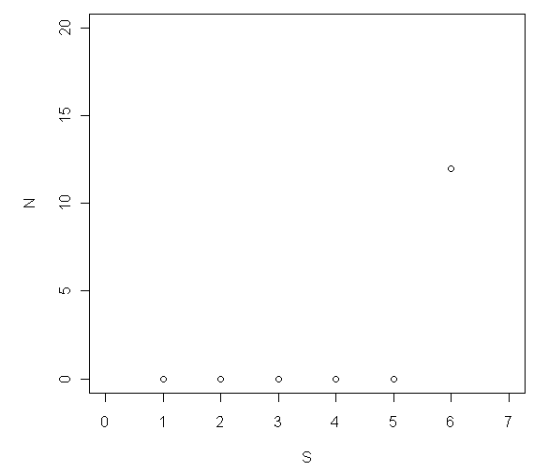
Distribution of patterns size N over time for CorpusP (a) N=1 (b) N=2 (c) N=12.



(a)



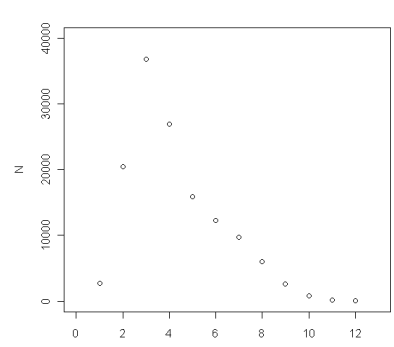
(b)



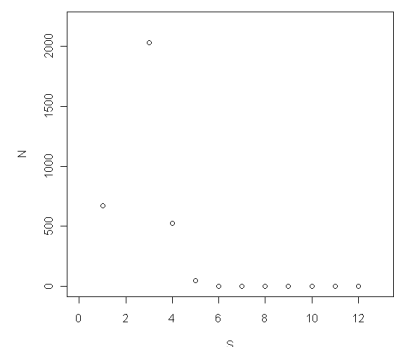
(c)

Experimental Results and Modeling with "Content-Words"

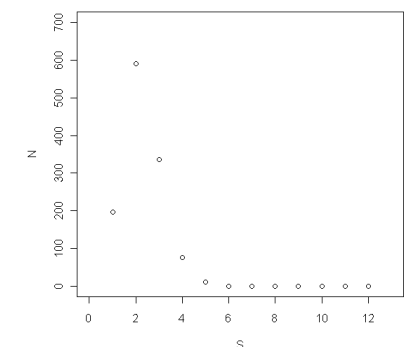
Distribution of patterns size given a period T for CorpusP (a) T=1 (b) T=2 (c) T=3 (d) T=4 (e) T=5 (f) T=6.



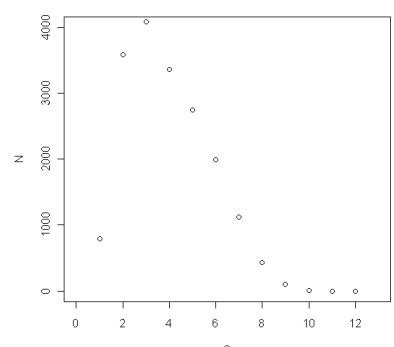
(a)



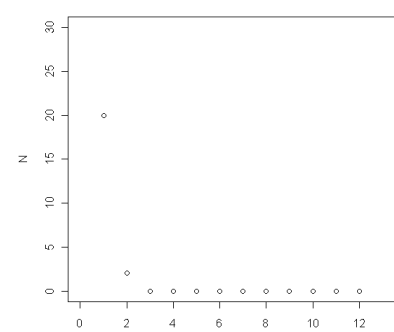
(b)



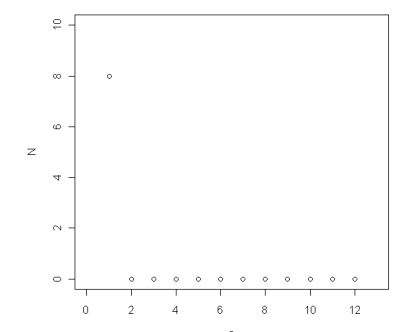
(c)



(d)



(e)



(f)

Experimental Results and Modeling with “Content-Words”

Distribution is shaped as local maximum distributions in the same way observed for author names but shape seems not exactly the same if we look at interval $T=1$ and $T=4$.

Distribution is enriched by a “bump” towards tail of distribution.

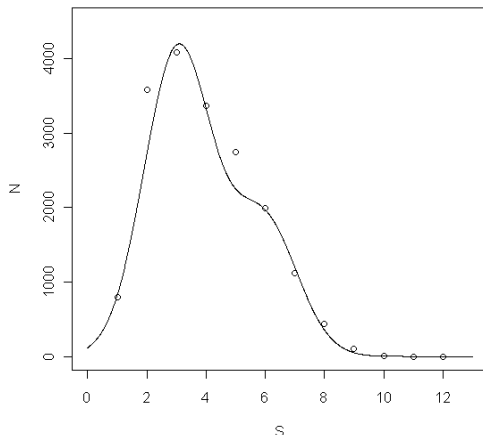
$$y = A \left(\alpha e^{-\frac{(x-\mu_1)^2}{2\sigma}} + (1-\alpha)\beta_t e^{-\frac{(x-\mu_2)^2}{2\sigma}} \right)$$

$$\text{with } \alpha_t = \{0,1\} \text{ and } \mu_2 \cong 2.\mu_1$$

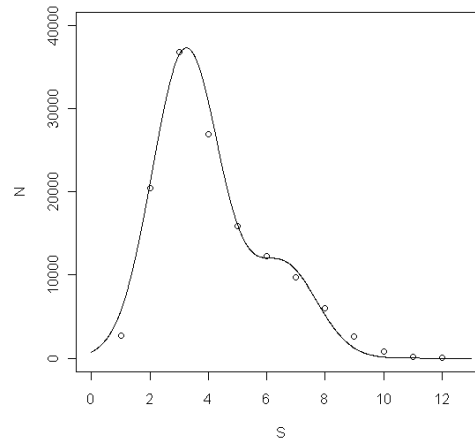
We modeled the asymmetric distribution with a mixed distribution composed by two weighted normal distributions.

Experimental Results and Modeling with “Content-Words”

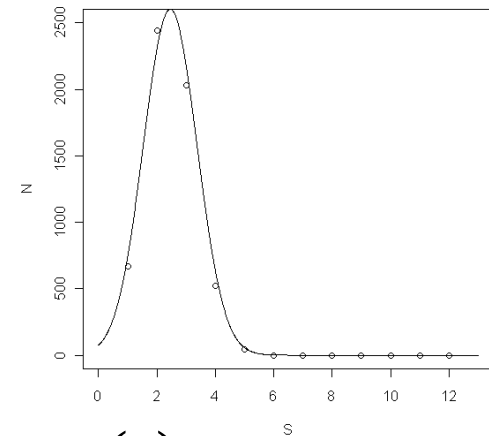
Modelling of Distribution of patterns according periot T for CorpusP (a) T=4 (b) T=1 (c) T=2 .



(a)



(b)

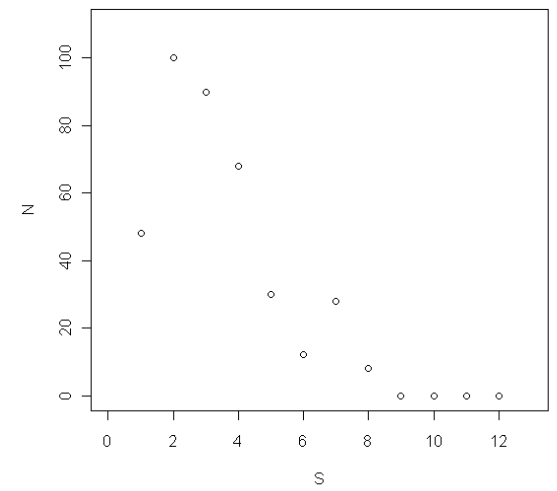
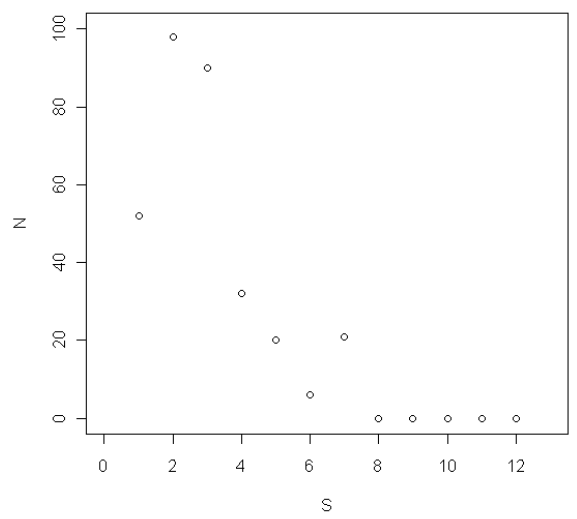


(c)

Using the library mixtool (R-Project, 2004) for learning parameters we find the following values: for T=1 at=0.77, m1=3.23, m2=6.56, s=1.14. For T=4 parameters are at=0.69, m1=3.04, m2=5.93, s=1.15 ; for T=2 at=1, m1=2.45, s=0.92. Weight at gets a binary domain value.

Experimental Results and Modeling with "Content-Words"

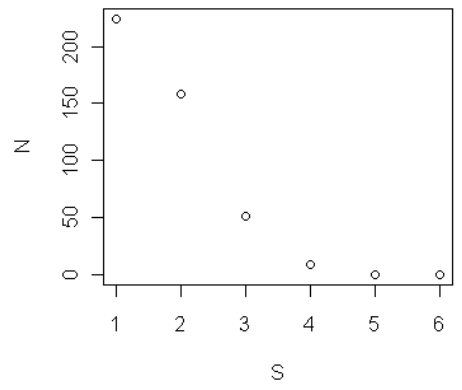
Distribution of maximal patterns by size given a period T for CorpusP (a) T=1 (b) T=4 .



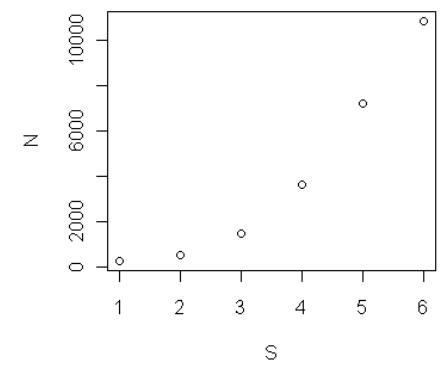
Let call a maximal pattern a pattern that does not belong to a pattern of larger size

Experimental Results and Modeling with "Content-Words"

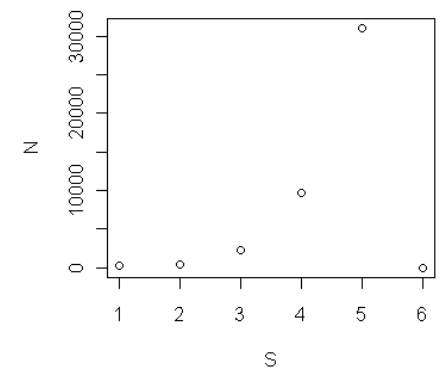
Distribution of patterns by size given a period T for CorpusE (a) T=2 (b) T=3 (c) T=4



(a)



(b)

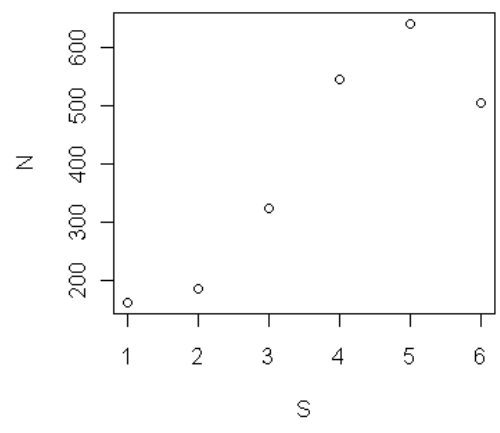


(c)

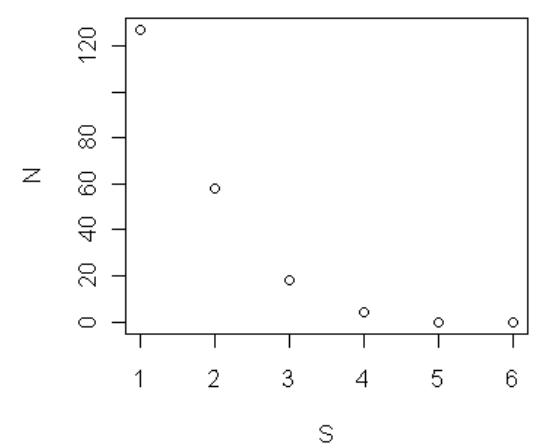
distribution of patterns by size and let us to understand that average can vary from a period to another but distribution can be modeled easily with $at=1$.

Experimental Results and Modeling with “Content-Words”

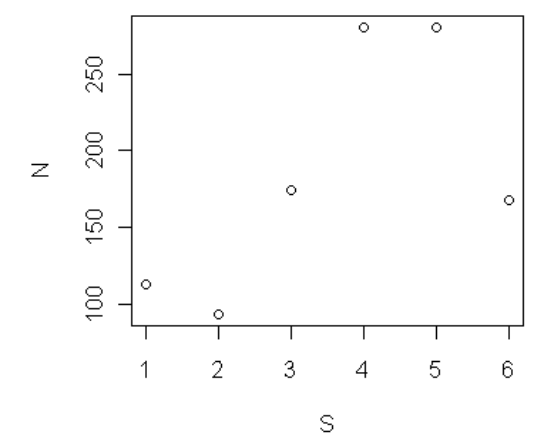
Distribution of patterns by size given a period T for CorpusC (a) T=1 (b) T=2 (c) T=3



(a)



(b)



(c)

Conclusions

Working hypotheses

- Processing **domain corpora**, in our case here biology,
- Analyzing distribution of “**content-words**”
- **Diachronic** phenomena.
- **Noun-phrases**, using different tools of named entities extraction. (named entity = a protein or a gene name)
- Extraction was achieved from **different text collections**, mouse embryo development, human embryo cell proliferation, prion, epidemiology and clustering techniques.

What we found

- About *Single “Content-Words”* a **linear-shape model** explains the regular presence of noun-phrases per sentence.
- As single noun-phrases do not provide real semantic information about the content of a corpus, we found that distribution of association size over time can be, sometimes but not all the times, a **mixed distribution** of small and double-size associations.

Perspectives

- Stable distributions require a **hypotheses framework**, some further studies need to be driven about contexts of distributions to embed them within more theoretical hypotheses as in synergetics framework

-
-

Special thanks to colleagues from INRA

Dr. Isabelle Hue - senior research scientist in molecular biology
for providing CorpusM and CorpusH

Dr. Marc Barbier - senior research scientist in sociology of science
for providing CorpusP and CorpusE