Entropic characterisation of termino-conceptual structure: A preliminary study

Kyo Kageura[1,2] Long-Huei Chen[1]
(1) Interfaculty Initiative in Information Studies, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113–0033, Japan
(2) Graduate School of Education, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113–0033, Japan
longhuei@g.ecc.u-tokyo.ac.jp, kyo@p.u-tokyo.ac.jp

RÉSUMÉ

Caractérisation entropique de la structure termino-conceptuelle : Une enquête préliminaire

Les termes représentent des concepts, qui consistent en des caractéristiques conceptuelles. Dans la formation sur le terrain du concept et du terme, qui est effectuée par les chercheurs, le processus est inversé : les éléments / caractéristiques conceptuels sont consolidés pour former des concepts, qui seront représentés par des termes. Les concepts n'existant pas a priori, ce processus est échafaudé par ce que nous pouvons appeler un "système termino-conceptuel". Les terminologues, tant dans la pratique que dans la recherche, ne font pas que cueillir et énumérer des termes; en plus ils analysent, décrivent et définissent les termes tout en systématisant les terminologies. Pour mener à bien ces tâches, les terminologues doivent se référer aux systèmes conceptuels, dans la mesure où ils contribuent à systématiser les terminologies ; les terminologues abordent donc également le domaine du système terminologique-conceptuel. Dans cet article nous appuyons le statut du domaine terminologique-conceptuel en proposant un procédé pour caractériser la structure du système terminologique-conceptuel en termes d'entropie. Nous analysons l'entropie des terminologies en langue anglaise de six domaines : l'agriculture, la botanique, la chimie, l'informatique, la physique et la psychologie.

ABSTRACT

Terms represent concepts, which consist of conceptual characteristics. In actual concept-term formation, which is done by researchers, the process is in reverse: conceptual elements/characteristics are consolidated to form concepts, which are represented by terms. As concepts do not exist on the fly, what we may call termino-conceptual system provides scaffolding in this process. Terminologists, both in practice and in research, do not only collect and list terms but also analyse, describe and define terms and systematise terminologies. To carry out these tasks, terminologists must refer to conceptual systems, to the extent that they contribute to systematising terminologies; terminologists thus also deal with the sphere of termino-conceptual system. In this paper, we consolidate the status of termino-conceptual sphere and propose a way to characterise the structure of termino-conceptual system by using entropy. The entropic characterisation of English terminologies of six domain, i.e. agriculture, botany, chemistry, computer science, physics and psychology are presented.

MOTS-CLÉS : Terminologie, entropie, structure de spécification, structure de classification.

KEYWORDS: Terminology, entropy, specification structure, classificatory structure.

1 Introduction

In societies of which technologies and specialised knowledge are inherent part, understanding of specialised concepts and proper treatment of technical terms are required in many social activities. Terminologists, both in practice and in research, play an important role in the activities that involve technical communication, including technical translation. The task of terminologists are not only to collect terms but also to analyse, describe and define terms and to systematise terminologies. In order to carry out these tasks, terminologists refer to conceptual systems that underline terminologies.

Concepts of a domain are formed and named as terms by researchers of the domain. Concepts consist of characteristics, and concept formation can be approximately regarded as structurisation of conceptual characteristics. This process is bound by existing conceptual system and terminology (Sager, 1990; Kageura, 2002). As terminological naming tends towards "transparency and consistency" (Sager, 1990:57), representational structure of terminologies reflect conceptual system to a substantial extent. We can identify here a system consisting of conceptual characteristics and concepts not in their abstract existence but as represented by terminologies, which we may call a sphere of "termino-conceptual" system.

This sphere operates implicitly as scaffolding for researchers in their activities, while it constitutes, in a sense, the main target of terminological work, as terminologists are concerned with terms and concepts to the extent that they are relevant to terminology processing and management. Nevertheless, this sphere *itself* has not been the explicit reflective target of terminological research (Kageura, 2015). Against this backdrop, this paper tries to consolidate the sphere of termino-conceptual system, introduces a simple information-theoretic approach to characterise the termino-conceptual structure, and analyses and characterises English terminologies of six domains, i.e. agriculture, botany, chemistry, computer science, physics and psychology.

2 An approximation to termino-conceptual sphere

2.1 Framework and layers for observing termino-conceptual structure

What we call here the sphere of termino-conceptual system is, as briefly stated in Introduction, the sphere where terminological representations and conceptual system meets. It contains conceptual characteristics, concepts and conceptual system to the extent that they are relevant to terms. Researchers' thinking process and external factors that lead them to come up with new concepts are out of this sphere. It contains terms and terminological system to the extent that they represent concepts. Linguistic features or social status of terminological elements are not a part of this sphere. While this sphere is defined as a theoretical abstraction, practical terminological work is carried out in or around this sphere. For instance, definitions given in terminological lexicons describe concepts only to the extent that they are identified within the terminological system that represent them.

Terms represent concepts, which consist of conceptual characteristics. Terms are "motivated" (de Saussure, 1911) to a substantial degree, representing important characteristics of the concepts by constituent elements of complex terms. For instance, "brain tumour" represents the concept of "the growth of abnormal cells (tumour) in the tissues of the brain"[1]. Two conceptual characteristics, i.e.

[1]National Cancer Institute. NCI Dictionary of Cancer Terms. https://www.cancer.gov/publications/dictionaries/cancer-

tumour as the genus concept and brain, the body part, as the core differentiating characteristic, are represented as nucleus and determinant of this term, respectively. A substantial amount – around 80 percent – of terms in most domains in many languages are complex (Cerbah, 2000; Nomura & Ishii, 1988). These complex terms represent important conceptual characteristics while at the same time showing the relative position of concepts represented by terms within the conceptual system (Sager, 1990; Kageura, 2012).

As a first step, therefore, we can approximate the termino-conceptual system by means of the surface structure of terminologies, under some simplifying assumptions: (a) constituent elements of terms represent conceptual characteristics without ambiguities; (b) terms represent important conceptual characteristics by their constituent elements; and (c) the structure of complex terms reflect the position, i.e. the nucleus and the determinants, of conceptual characteristics. We basically adopt here that terms/concepts are formed by combining constituent elements/conceptual characteristics. This approximation can enable us to observe differences in linguistic representations of concepts in different domains or in different languages.

Under the assumptions adopted here, we can define layers of characterising termino-conceptual structure, starting from a set of conceptual characteristics (cf. Sager, 1990; Kageura, 2002):

1. Each subject domain has a set of conceptual characteristics, by using which concepts are consolidated. We call it the *base set*.
2. These conceptual characteristics are used repeatedly in forming concepts; this assigns distribution to conceptual characteristics in the domain. We call this layer the *selection structure*.
3. These conceptual characteristics are used in combination with other constituent elements to form concepts, either as a nucleus of concepts, fixing the core of the concepts, or as determinants, to specify the concepts. We call this layer the *specification structure*.
4. These conceptual characteristics are used to form concepts in such a way that the paradigmatic or classificatory positions of the concepts within the conceptual system are reflected. We call this layer *classificatory structure*.

Note that these layers are not for describing or analysing given terms or terminologies, but for characterising the termino-conceptual sphere that underlies given terms and terminologies, from the point of view of the organisation of conceptual characteristics. We can regard the status of termino-conceptual sphere to given terms and terminologies as somewhat analogous to the status of "language model" to given corpus data in NLP, although the compositional elements and their arrangements of the models and the theoretical status are different[2].

2.2 Entropic characterisation

Assuming one-to-one correspondence between conceptual characteristics and constituent elements of terms, we can characterise these layers by analysing given terminologies (hence we also use elements to refer to conceptual characteristics). In this study, as a first step, we characterise the

terms/def/brain-tumor. Accessed 31 March 2019

[2]Unlike language models, the descriptions here cannot be directly used for generating or processing possible terms. Also, terminological models need to address "realistic possibility of existence" (Kageura, 2002) while language models in general deal with any sort of reasonably "well-formed" expressions.

structure of these layers by using entropy (Cover & Thomas, 2006; Alajaji & Chen, 2018). Let X be a discrete random variable taking values in the finite base set of constituent elements/conceptual characteristics $\mathcal{X}$, to which probability distribution is given. Then the entropy $H(X)$ of X is:

$$H(X) = -\sum_{i=1}^{|\mathcal{X}|} P(x_i) \log_2 P(x_i).$$

The remaining issue is to define and estimate a probability distribution of elements at each layer. By definition, elements are distributed uniformly in the *base set* layer. This provides a point of reference to the structure of other layers, as entropy takes its maximum for the *base set* layer. We can assign probability distribution to the layer of *selection structure* by giving sample relative frequencies of elements in given terminological data[3]. For the layer of *specification structure*, we define the head/nucleus-modifier/determinant[4] directed graph from terminologies (henceforth *specification graph*), which enables us to analyse the overall system of elements representing specification structure. For the layer of *classificatory structure*, we construct bibliographic coupling (Kessler, 1963) and co-citation (Small, 1973) graphs, both undirected, from the specification graph, by adding edges among the elements that modify/are modified by the same element, respectively. In other words, they show how systematic a set of modifiers classify the concept represented by heads (bibliographic coupling) and how systematic a set of heads undergo a characterisation by modifiers (thus we call them *classification graph* hereinafter)."

Figure 1 shows the specification graph (left) and classification graph made by bibliographic coupling (right) constructed from a small artificial terminology consisting of 12 terms, i.e. “text classification”, “document classification”, “information retrieval”, “library classification”, “automatic text classification”, “document information”, “medical information”, “medical aid system”, “medical diagnosis”, “disease diagnosis”, “diagnosis record”, “disease image record”.

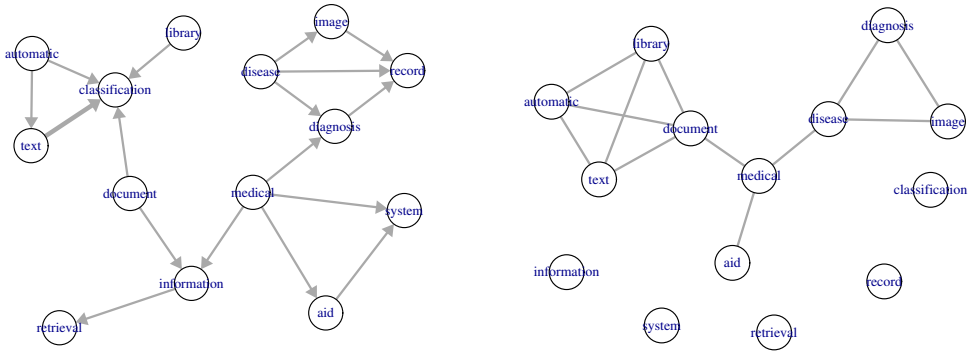

Figure 1: Specification graph (left) and Classification graph (right)

Once we define a graph, a number of methods to estimate graph entropy become immediately available (Dehmer, 2008b; Dehmer & Emmert-Streib, 2008; Dehmer, 2008a; Dehmer & Emmert-Streib,

[3]A simple estimation of probabilities of elements by sample relative frequency ignores the existence of potential elements that do not occur in given terminological data. The issue starts from defining the *base set*. For entropy estimation in the present scheme, we have several choices: (i) temporarily ignore this issue; (ii) estimate the unseen elements first (Baayen, 2001; Kageura, 2012), assign probabilities to them and measure entropies; (iii) theorise the change in entropy values in accordance with the size of the data and extrapolate it to obtain entropy rate (Takahira *et al.*, 2016), and (iv) use methods to deal with this issue in the process of entropy estimation (Hausser & Strimmer, 2009; Archer *et al.*, 2014). At this stage we took the first approach, for this is sufficient to contrast structures in different layers and of different domains.

[4]We use “head” and “modifier” to refer to constituent elements of terms and “nucleus” and “determinants” to refer to conceptual positions but within the immediate context of this paper we use them interchangeably.

2009; Dehmer, 2011; Dehmer & Mowshowitz, 2011; Körner, 1973; Mowshowitz, 1968; Mowshowitz & Dehmer, 2012; Simonyi, 1995; Trucco, 1956). We have to resist the temptation to resort to complex graph entropic measures[5]. We introduce the graph to characterise the layers of specification and classificatory structures, so gaining due probability distributions over constituent elements which we can make sense of as characterising these layers is the goal.

Upon examining possible approaches, including the way to weight vertices in the graph (Bonacich, 1987; Estrada, 2015, 2016; Kolaczyk, 2009; Newman, 2010), we adopted PageRank (Avrachenkov *et al.*, 2015; Langville & Meyer, 2012; Page *et al.*, 1998), as the algorithm assigns a probability to each vertex and these probabilities can be interpreted in terms of a random walk on the graph, i.e. the probability of arriving at a vertex in the graph. To the extent that the graphs defined above reflect, *en masse*, the specification and classificatory structures, we can interpret these probabilities as the likelihood of deploying elements in concept formation, captured at these layers. More specifically:

- Probability distribution of elements defined on a specification graph reflects the tendencies of conceptual characteristics being used as nuclei in terminologies[6] (nucleus-oriented specification graph);
- Probability distribution defined on a specification graph with directions reversed reflects the tendencies of characteristics being used as determinants (determinant-oriented specification graph);
- Probability distribution defined on a classification graph constructed by bibliographic coupling reflects the tendency of elements to be deployed as "sibling" characteristics in modifiers/determinants (determinant-oriented classification graph);
- Probability distribution defined on a classification graph constructed by co-citation reflects the tendency of elements to be deployed as sibling characteristics in heads/nuclei (nucleus-oriented classification graph).

Entropies can be measured straightforwardly for each of these probability distributions.

3 Analysis of English terminologies of six domains

Here we analyse and describe the termino-conceptual structures of English terminologies, based on the theoretical and methodological framework defined in Section 2.

3.1 Terminological data

Within the framework defined above, termino-conceptual structure of terminologies of six domains are analysed, i.e. agriculture (Agr) (Japanese Ministry of Education, 1986a), botany (Bot) (Japanese Ministry of Education, 1986b), chemistry (Chm) (Japanese Ministry of Education, 1986c), computer science (Cmp) (Aiso, 1993), physics (Phy) (Japanese Ministry of Education, 1990b), and psychology (Psy) (Japanese Ministry of Education, 1986d). To improve the correspondence between constituent

[5]Incidentally it is pointed out that the graph entropic measures need to be used with care (Zenil *et al.*, 2017).

[6]This contrasts with observing distributions of constituent elements used as heads and modifiers of terms in terminologies.

Dom	T	N	V	N/T	N/V	N_c	V_c	N_c/T	N_c/V_c
Agr	15614	30436	8409	1.95	3.62	29519	7087	1.89	4.17
Bot	9774	16152	7032	1.65	2.30	15992	6409	1.64	2.50
Chm	10912	19110	6795	1.75	2.81	18934	5801	1.74	3.26
Cmp	14099	32410	4932	2.30	6.57	31770	3938	2.25	8.07
Phy	9974	21191	4914	2.12	4.31	20718	4267	2.08	4.86
Psy	5866	11477	4026	1.96	2.85	11068	3467	1.89	3.19

Table 1: Basic quantities of the terminologies of six domains

elements and conceptual characteristics, we removed 20 function words[7] and applied stemming, using porter2 algorithm (Dempsey, 2016). Table 1 gives the basic quantities of these terminologies[8]. In Table 1, T stands for number of terms, N and V the number of constituent elements in token and in type, respectively, and N_c and V_c are the number of constituent elements in tokens and in types after functional elements were removed.

3.2 Specification and classification graphs

Based on the data, specification and classification graphs are constructed as follows[9]:

1. Head-modifier relations are automatically extracted by simple rules that reflect the fact that English complex nouns are head final and nominal phrases with prepositions are head initial. For complex terms consisting of more than three elements, all the possible head-modifier relations are used. For instance, “abnormal grain growth”, “abnormal-grain”, “abnormal-growth”, and “grain-growth” are extracted as head-modifier relations. We took into account the reverse modification relations for phrasal terms. For instance, the head-modifier relation for the “abandonment of cultivation” is identified as “cultivation-abandonment”.
2. The directed graphs for “head $\leftarrow$ modifier” (nucleus-oriented) and for “modifier $\leftarrow$ head” (determinant oriented) are constructed as shown in Figure 1.
3. Bibliographic coupling and co-citation graphs are constructed based on the head-modifier specification graph. When two elements are in specification relations, i.e. directly connected in the specification graph, we did not link them in the classification graphs.

Basic information of the specification graphs are given in Table 2[10]. The graphs are not connected in the terminologies, but there is a single giant component in each terminology. We will explain how random-walk probabilities are given in 3.4.

[7]We started from a larger list of English stopwords used in IR (http://www.lextek.com/manuals/onix/stopwords1.html) and defined 20 function words for removal.

[8]The original data are the same as those used in (Asaishi & Kageura, 2011), but the statistics are slightly different, due to normalisation including typo corrections and different ways of pre-processing.

[9]We used R package of igraph for graph construction and processing (Csardi & Nepusz, 2006).

[10]V stands for vertex, E stands for edge, Comp. stands for components, D is diameter and PL is average path length. Type and token for isolates are different due to stemming. Information for classification graphs are not given due to space limitations.

	Isolated V		Connected V		No. of	No. of	Maximum Component			
Dom	Type	Token	Type	Token	Edges	Comp.	V	E	D	PL
Agr	2442	2489	4645	27030	14585	68	4506	14507	17	4.69
Bot	3137	3213	3272	12799	6389	68	3106	6282	23	5.66
Chm	2360	2399	3441	16535	8308	74	3290	8229	18	5.18
Cmp	706	719	3232	31051	17268	27	3169	17210	21	3.84
Phy	733	736	3534	19982	11368	56	3400	11279	13	5.61
Psy	951	957	2516	11423	5575	72	2322	5401	19	5.75

Table 2: Basic quantities of the specification graphs of six domains

3.3 Specification and classification graphs

To interpret entropies measured for these graphs, we introduced Erdös-Rényi (ER) random graph (Erdös & Rényi, 1959; Frieze & Karoński, 2015), Barabási-Albert (BA) preferential attachment graph (Barabási & Albert, 1999), and Watts-Strogatz (WS) small world graph (Watts & Strogatz, 1998), which are used frequently as points of reference to analyse the nature of real-world graphs (Estrada, 2015, 2016; Kolaczyk, 2009; Newman, 2010):

- We constructed an ER directed graph with the same order (the number of vertices) and size (the number of edges) as each of the terminologies. This provide the baseline in which the relationships between nuclei and determinants are randomly chosen. As ER graph is "reversible", it is used as a point of reference for both nucleus-oriented and determinant-oriented specification graphs. Thus ER graphs give reference situations in which modifying and modified relations are "symmetric." The same holds for WS graph.
- For BA graph, we adopt the same order and size as each of the terminologies, adding at each time step the number of edges given by the integer division of size by order. We used linear preferential attachment, and allowed multiple edges. BA model is known to generate degree distribution that follows power law with the power approaching -3 (Kolaczyk, 2009; Newman, 2010). As degree distributions correlates with the frequency distribution of constituent elements and it is known that constituent elements roughly follow power law (Kageura, 2012), BA model gives a point of reference to evaluate specification graphs in the light of degree distributions. BA graphs are not reversible, so we use two BA graphs for each terminology.
- For WS graph, we first constructed a directed lattice graph with the same order and size as each of the terminologies, connecting "neighbouring" lattices by the number of edges given by the integer division of size by order of the terminologies, and then rewired the edges with rewire probabilities as 0.01, 0.05, 0.1 and 0.2. We report here the entropies for WS graphs with rewire probabilities as 0.2, because they give on average values in diameter and average path length closest to those of the largest components of the corresponding specification graphs. WS graph is essentially reversible, so we use only one graph. WS graph gives a point of reference in which the use of conceptual characteristics in determinants and nuclei start from uniform distributions and then some motivated formations are introduced.

As these graphs statistically differ each time they are generated with the same parameters, we generated each graph 100 times and took the probability distributions based on PageRank, mean of entropies and other basic measures. Note that BA and WS graphs are connected, while ER graphs may

Dom	Base	Selection	Specification		Classificatory	
			nucleus	determinant	nucleus	determinant
Agr	8.866	7.872 (0.887)	7.084 (0.799)	7.820 (0.882)	7.056 (0.796)	7.497 (0.846)
Bot	8.766	7.991 (0.912)	6.998 (0.799)	6.777 (0.773)	6.685 (0.763)	7.417 (0.846)
Chm	8.666	7.751 (0.894)	6.765 (0.781)	6.512 (0.751)	6.718 (0.775)	7.386 (0.852)
Cmp	8.278	6.986 (0.844)	6.400 (0.773)	7.105 (0.858)	6.577 (0.794)	7.153 (0.864)
Phy	8.359	7.327 (0.877)	6.672 (0.798)	6.539 (0.782)	6.586 (0.788)	7.433 (0.889)
Psy	8.151	7.285 (0.894)	5.551 (0.681)	6.820 (0.837)	6.352 (0.779)	7.058 (0.866)

Table 3: Entropies of different layers of termino-conceptual structure

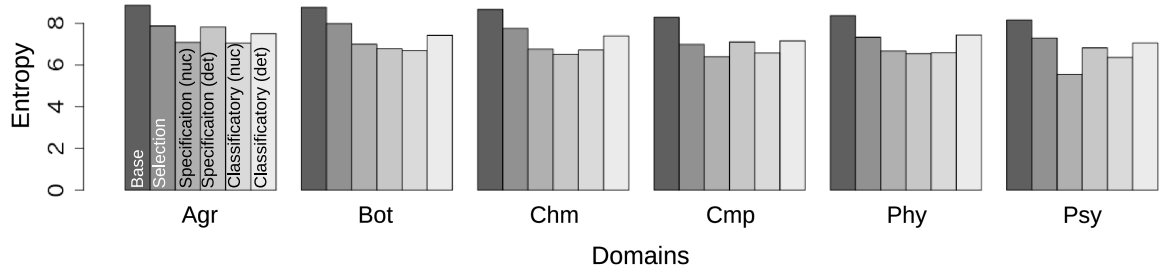

Figure 2: Entropies of different layers of termino-conceptual structure

not necessarily be connected. Classification graphs are generated from these specification graphs. They may not be connected.

3.4 Entropic nature of termino-conceptual sphere

Based on this setup, we measured entropies at each layer of termino-conceptual structure[11]. For the entropies of the selection structure, we used maximum likelihood estimators[12]. For specification and classificatory graph, we adopted the following setups:

- In applying PageRank, the restart probability, used to avoid walks to end at sinks, was set to 0.01, i.e. damping factor was set to 0.99. This is lower than what is widely used (0.15).
- The probabilities assigned to the vertices of each component were normalised by the number of element tokens to make the interpretation intuitively natural[13]

Table 3 shows the entropies of terminologies (ratio to base set entropies are given in brackets). To facilitate the analysis, Figure 2 gives the corresponding barplot. As entropies measured by MLE estimators depend on the number of items, it is not theoretically simple to directly compare the nature of the termino-conceptual structure across domains. We observe several tendencies. First, specification layers and classification layers are generally more "structured" than selection layer. Exceptions are determinant-oriented specification and classificatory structures (fourth and the rightmost bars in Figure 2, respectively) in computer science, and to some extent agriculture. This indicates that in these domains, conceptual characteristics are deployed for determining concepts more evenly than

[11]We used R package of entropy for entropy calculation (Hausser & Strimmer, 2009).

[12]We also observed shrinkage estimator (Hausser & Strimmer, 2009) for selection structure but the values differ little in our data. So we report MLE values.

[13]This normalisation in reality made little difference in entropies.

the other domains. Secondly, in agriculture, computer science and psychology, entropies for nuclei-oriented specification graphs are smaller than those for determinant-oriented specification graphs, while these are the other way round in botany, chemistry and physics. Third, determinant-oriented classificatory sphere is less structured and in most domains close to the entropies of the selection layer. This may imply that conceptual characteristics used to differentiate nuclei are systematic, in the sense that different nuclei are specified in similar manner.

Tables 4 and 5 show entropies for ER, BA (nucleus-oriented and determinant-oriented) and WS specification graphs. Table 4 also gives diameter (D) and average path length (PL). Figure 3 shows the barplot of the entropies of the six domains, with the bars from left to right showing: nucleus-oriented specification entropy of terminology, determinant-oriented specification entropy of terminology, entropy of ER specification graph, nucleus-oriented entropy of BA specification graph, determinant-oriented entropy of BA specification graph, entropy of WS specification graph. Figure 4 shows the barplot of the entropies of the classificatory structure, with the bars from left to right showing the entropies of the corresponding classificatory graphs and reference graphs. These values are the means of 100 graphs each. We do not give standard deviations due to space limitations. Note that ER and WS are very close in entropy values.

	ER			BA				WS		
Dom	Entr.	D	PL	"Nuc"	"Det"	D	PL	Entr.	D	PL
Agr	8.156	17.0	7.4	5.645	8.354	11.4	3.1	8.152	17.2	7.4
Bot	7.719	28.9	11.0	5.488	8.017	10.8	2.0	7.709	28.3	11.0
Chm	7.794	22.1	8.9	5.498	8.067	11.0	3.0	7.786	21.8	8.9
Cmp	7.931	10.1	5.0	5.341	7.980	11.0	3.0	7.932	10.1	5.0
Phy	7.890	16.3	7.0	5.500	8.082	11.0	3.0	7.887	16.2	7.0
Psy	7.466	23.4	9.3	5.288	7.754	10.2	2.8	7.460	23.2	9.3

Table 4: Entropy, diameter and average path lengths of reference specification graphs

	ER	BA		WS		ER	BA		WS
Dom		"Nuc"	"Det"		Dom		"Nuc"	"Det"	
Agr	8.230	6.573	8.081	8.229	Cmp	7.969	6.350	8.026	7.969
Bot	7.721	6.421	7.403	7.721	Phy	7.963	6.330	7.850	7.962
Chm	7.851	6.266	7.599	7.850	Psy	7.507	5.994	7.288	7.508

Table 5: Entropy, diameter and average path lengths of reference classification graphs

From Figure 3, we can observe that, in all the domains, entropies of both nucleus-oriented and determinant-oriented specification structures of terminologies are smaller than the entropies of ER, determinant-oriented BA and WS graphs. Terminologies are more structured than these reference graphs. On the other hand, entropies of nucleus-oriented specification structures are larger than the entropies of nucleus-oriented BA graph. This indicates that tendencies for preferential combinations in terminologies are weaker than the theoretical BA model we adopted here. It is notable that in psychology the difference is much smaller compared to other domains. From Figure 4, we can observe corresponding tendencies for the entropies of classificatory structures, with the apparent differences between terminologies and BA graphs being smaller. All in all these show that specification and classification structures of terminologies may be modelled starting from BA preferential-attachment graph. This has been informally indicated by the fact that constituent elements follow power law. The preference is weaker in general in terminologies, though.

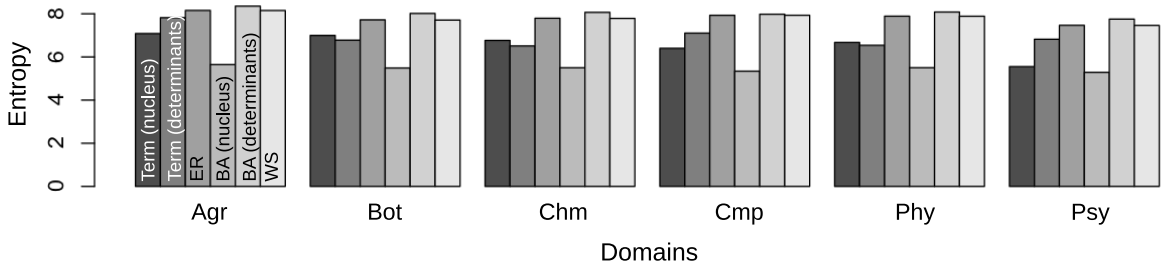

Figure 3: Entropies of specification structure and of reference graphs

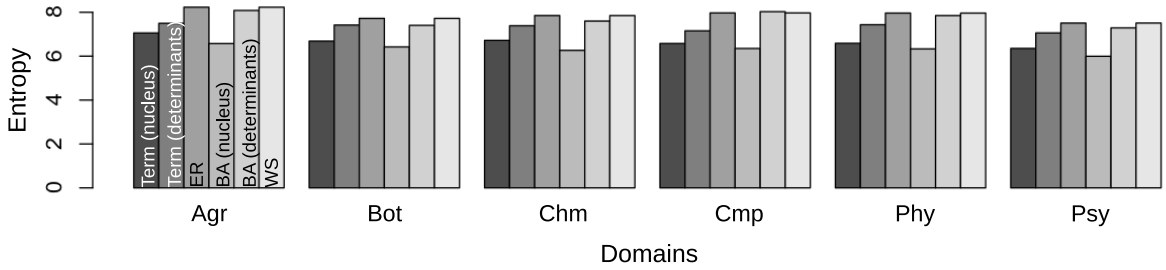

Figure 4: Entropies of classificatory structure and of reference graphs

4 Conclusions and outlook

In this paper, we carried out three tasks, i.e. (1) consolidated the theoretical object which can be called termino-conceptual sphere and argued that terminologists essentially work on this area; (2) defined layers relevant to concept/term formation and introduced ways to characterising these layers; and (3) described the structure of termino-conceptual sphere for the terminologies of the six domain. Although the structures of terminologies became visible, the descriptions given in this paper serve for understanding only the overall structural nature of terminologies.

Several issues remain and indicate directions for further research, of which we mention three here. Extending the viewpoints of observations by introducing different summary measures and more analytical features, within the framework we defined here is the first theoretical task. Graph-oriented measures are abundant; but the give proper interpretation is a different issue. We also intend to carry out the descriptive studies of terminologies of different domains and of different languages.

Second, when taking about concepts and conceptual characteristics, we assumed one-to-one correspondence between conceptual characteristics and constituent elements of terms. We did not use, for instance, distributional representations. In this paper, this was a deliberate decision. In terminology, exactness in forms is important as in proper names; one cannot replace a constituent element of terms with another element with similar meanings. As such, symbolic identity bear more information than semantic similarity. This, however, is only one side of the story. To the extent that terms use linguistic items as their representational elements, terms, concepts and their relations have certain degree of flexibility (Rey, 1995). To fully capture the nature of terminological structure, it would be beneficial to use, for instance, distributional representations of concepts.

Third, we can explore how to use the information contained in these layers for automatic terminology processing. From this point of view, one possibility is to construct distributional representation of concepts for constituent elements by using the information contained in these layers. We may further use the information in the layers to improve terminology augmentation in generate-and-validate framework (Iwai *et al.*, 2016) and to improve performance of term translations.

References

AISO H. (1993). *Terminological Lexicon of Information Processing*. Tokyo: Ohm.

ALAJAJI F. & CHEN P.-N. (2018). *An Introduction to Single-User Information Theory*. Singapore: Springer.

ARCHER E., PARK I. M. & PILLOW J. W. (2014). Bayesian entropy estimation for countable discrete distributions. *Journal of Machine Learning Research*, **15**, 2833–2868.

ASAISHI T. & KAGEURA K. (2011). Comparative analysis of the motivatedness structure of Japanese and English terminologies. In *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence (TIA 2011)*, p. 38–44.

AVRACHENKOV K., KADAVANKANDY A., PROKHORENKOVA L. O. & RAIGORODSKII A. (2015). *PageRank in Undirected Random Graphs*. Technical report, INRIA.

BAAYEN R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

BARABÁSI A. L. & ALBERT R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–512.

BONACICH P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, **92**, 1170–1182.

CERBAH F. (2000). Exogeneous and endogeneous approaches to semantic categorization of unknown technical terms. In *COLING 2000*, p. 145–151.

COVER T. M. & THOMAS J. A. (2006). *Elements of Information Theory*. New York: John Wiley & Sons.

CSARDI G. & NEPUSZ T. (2006). The igraph software package for complex network research. *International Journal of Complex Systems*, p. 1695.

DE SAUSSURE F. (1911). *Linguistique Générale 1910–1911 (Lecture notes taken by Emile Constantin). X vols.* Genève: Université de Genève.

DEHMER M. (2008a). Information processing in complex networks: Graph entropy and information functionals. *Applied Mathematics and Computation*, **201**, 82–94.

DEHMER M. (2008b). A novel method for measuring the structural information content of networks. *Cybernetics and Systems: An International Journal*, **39**, 825–842.

DEHMER M. (2011). Information theory of networks. *Symmetry*, **3**, 767–779.

DEHMER M. & EMMERT-STREIB F. (2008). Structural information content of networks: Graph entropy based on local vertex functionals. *Computational Biology and Chemistry*, **32**, 131–138.

M. DEHMER & F. EMMERT-STREIB, Eds. (2009). *Analysis of Complex Networks: From Biology to Linguistics*. New York: John Wiley & Sons.

DEHMER M. & MOWSHOWITZ A. (2011). A history of graph entropy measures. *Information Sciences*, **181**, 57–78.

DEMPSEY E. (2016). *Porter2 Stemmer Documentation: Release 1.0*. Technical report, https://readthedocs.org/projects/porter2-stemmer/downloads/pdf/latest/.

ERDÖS P. & RÉNYI A. (1959). On random graphs. *Publicationes Mathematicae*, **6**, 290–297.

ESTRADA E. (2015). *A First Course in Network Theory*. Oxford: Oxford University Press.

ESTRADA E. (2016). *The Structure of Complex Networks: Theorpy and Applications*. Oxford: Oxford University Press.

FRIEZE A. & KAROŃSKI M. (2015). *Introduction to Random Graphs*. Cambridge: Cambridge University Press.

HAUSSER J. & STRIMMER K. (2009). Entropy inference and James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, **10**, 1469–1484.

IWAI M., TAKEUCHI K., ISHIBASHI K. & KAGEURA K. (2016). A method of augmenting bilingual terminology by taking advantage of the conceptual systematicity of terminologies. In *Computerm 2016*, p. 30–40.

JAPANESE MINISTRY OF EDUCATION (1986a). *Scientific Terms: Agriculture*. Tokyo: JSPS.

JAPANESE MINISTRY OF EDUCATION (1986b). *Scientific Terms: Botany*. Tokyo: Maruzen.

JAPANESE MINISTRY OF EDUCATION (1986c). *Scientific Terms: Chemistry*. Tokyo: Chemical Society of Japan.

JAPANESE MINISTRY OF EDUCATION (1986d). *Scientific Terms: Psychology*. Tokyo: JSPS.

JAPANESE MINISTRY OF EDUCATION (1990b). *Scientific Terms: Physics*. Tokyo: Baifukan, 2nd edition.

KAGEURA K. (2002). *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth*. Amsterdam: John Benjamins.

KAGEURA K. (2012). *The Quantitative Analyses of the Dynamics and Structure of Terminologies*. Amsterdam: John Benjamins.

KAGEURA K. (2015). Terminology and lexicography. In H. J. KOCKAERT & F. STEURS, Eds., *Handbook of Terminology*, p. 45–59. Amsterdam: John Benjamins.

KESSLER M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, **14(1)**, 10–25.

KOLACZYK E. D. (2009). *Statistical Analysis of Network Data*. New York: Springer.

KÖRNER J. (1973). Coding of an information source having ambiguous alphabet and the entropy of graphs. *Transactions of the Sixth Prague Conference on Information Theory*, p. 411–425.

LANGVILLE A. N. & MEYER C. D. (2012). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton: Princeton University Press.

MOWSHOWITZ A. (1968). Entropy and the complexity of the graphs: I. An index of the relative complexity of a graph. *Bulletin of Mathematical Biophysics*, **30**, 175–204.

MOWSHOWITZ A. & DEHMER M. (2012). Entropy and the complexity of graphs revisited. *Entropy*, **14**, 559–570.

NEWMAN M. E. J. (2010). *Networks: An Introduction*. Oxford: Oxford Universiy Press.

NOMURA M. & ISHII M. (1988). *A List of Morpheme Combinations in Japanese Scientific Terms*. Technical report, National Language Research Institute.

PAGE L., BRIN S., MOTWANI R. & WINOGRAD T. (1998). *The PageRank Citation Ranking: Bringing Order to the Web*. Technical report, Computer Science Department, Stanford University.

REY A. (1995). *Essays on Terminology*. Amsterdam: John Benjamins.

SAGER J. C. (1990). *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.

SIMONYI G. (1995). Graph entropy: A survey. In W. COOK, L. LOVÁSZ & P. SEYMOUR, Eds., *Combinatorial Optimization*, p. 399–441. Province: American Mathematical Society.

SMALL H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, **24(4)**, 265–269.

TAKAHIRA R., TANAKA-ISHII K. & DĘBOWSKI L. (2016). Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora. *Entropy*, **18**, 364.

TRUCCO E. (1956). A note on the information content of graphs. *Bulleting of Mathematical Biology*, **18(2)**, 129–135.

WATTS D. J. & STROGATZ S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

ZENIL H., KIANI N. A. & TEGNÉR J. (2017). Low algorithmic complexity entropy-deceiving graphs. *Physical Review E*, **96**.