Agris on-line Papers in Economics and Informatics

Volume VI Number 4, 2014

Quantitative Differences among Normal and Knowledge Texts on Agriculture Waste Processing

T. Horáková, M. Houška

Faculty of Economics and Management, Czech University of Life Sciences Prague, Czech Republic

Anotace

Cílem článku je identifikovat rozdíl mezi vzdělávacími texty psanými běžnou formou a texty znalostními, které byly vytvořeny záměrným použitím metod znalostního inženýrství. Výzkumný vzorek tvoří 60 dokumentů – vzdělávacích textů z oblasti zpracování zemědělských odpadů, které byly autory převedeny do znalostní podoby. Nad sadou indikátorů, které se používají pro hodnocení didaktických textů, byly formulovány pracovní a operační hypotézy, jejichž platnost byla testována pomocí párového t-testu. Ukázalo se, že znalostní forma vzdělávacích textů vykazuje statisticky významně (α = 0,05) nižší koeficient celkové obtížnosti, když je při srovnatelném množství faktických a technických informací složen z významně většího počtu jednoduchých vět spojených v souvětí reprezentující znalost. Na základě významně větší frekvence vybraných identifikátorů je pak možné oba typy texů odlišit i formálně, na čemž je možné založit další výzkum: automatizované rozpoznávání typu vzdělávacího textu a měření obsahu znalostí, které jsou v něm uvedeny.

Klíčová slova

Zemědělské vzdělávání, znalostní a informační systémy, zemědělské odpady, literární styl, znalostní jednotka, celková obtížnost textu.

Abstract

The objective of this work is to identify the differences among educational texts written in two styles: normal educational text and their knowledge form. The research sample consists of 60 documents – educational texts on agriculture waste processing – converted by the authors into the knowledge form. Over the set of indicators used for evaluating the educational texts, we formulated working and operational hypotheses and validated them using the paired sample t-test. The results show that the complex text difficulty rate of knowledge texts is significantly (α = 0.05) lower than of the normal texts. They present the same amount of information logically divided into more simple sentences merged to complex sentences. Based on the difference in frequencies of selected identifiers we are able to distinguish the literary styles. The further research aims at an automatic recognition of the text styles and measuring the amount of knowledge inside the text.

Key words

Agriculture education, knowledge and information systems, agriculture wastes, literary style, knowledge unit, complex text difficulty rate.

Introduction

Classification of literary styles of texts is a common issue solved by many researchers from more points of view. Cortina-Borja and Chappas (2006) quantified the literary style of various forms of media, including the new ones (broadsheet and tabloid newspapers, technical periodicals and television news scripts). It allowed them to investigate the richness of vocabulary exhibited in these texts under the proposition that the writing style usually varies depending on the targeted readership or audience. Graham et al. (2012) state that in literature, there is an established set of techniques that have been successfully leveraged in the statistical analysis of literary style, most often to answer questions of authenticity and attribution. In their work, they suggest that the progress made and statistical techniques developed in understanding the visual processing

as it relates to natural scenes can serve as a useful model and inspiration for visual stylometric analysis.

In connection with the analysis of the literary styles of the documents, there is another issue worth solving: how to measure (ideally through quantitative characteristics) information content of the documents. This issue is really important in education, because it can influence the learning outcomes of the educational process (D. Newton, L. Newton, 2009). Duric and Song (2012) or Asaishi (2011) dealt with the analysis of educational texts. The aspects that were evaluated and measured included, among others, the extent of having the textbook equipped from the didactics point of view, the extent of the difficulty of the text, the analysis of terms, the extent of the information density, and so on. Just these authors inspired us to carry out the research presented in this paper.

In education, the main focus is put on the transfer of knowledge. We feel the ability of measuring the knowledge content in educational texts or textbooks as one of the critical factors in evaluation of the quality of the textbooks. On the other hand, according to our best knowledge no such metrics for measuring the knowledge content in the text (knowledge density, number of pieces of knowledge, etc.) have been developed and published. The objective of this work is to compare quantitative indicators and parameters of two types of texts: normal text without any corrections, and knowledge text created using the methods of Knowledge Engineering. In particular, the knowledge unit as the representation of knowledge in natural language is used. When the differences in quantitative indicators are identified and described, we can formulate more advanced hypotheses on the influence of the indicators on measuring the knowledge content of the text as an input for further research.

In this work we continue in our research on determining the quantitative characteristics of normal and knowledge texts. Previously (Rauchová et al., 2014), we tested the further presented methodology (see Materials and methods) and anticipated the quantitative characteristics of the text, which could be of the largest potential to distinguish among the text types. As we found, it is worth dealing mainly (but not exclusively) with the following indicators (see Materials and methods for their definition):

- semantic difficulty rate;
- syntactic difficulty rate;
- complex text difficulty rate;
- technical and factual information per words;
- number of concepts.

Apart from the previous analysis on micro-samples of the texts (Rauchová et al., 2014), other authors (e.g. McCrory and Stylianides (2014) or Miller (2011)) support our arguments for choosing this set of the indicators as well. The objective of the current work is to use statistically significant samples of homogeneous texts on agriculture waste processing and prove or disprove the following working hypotheses:

H1.0: The complex text difficulty rate (T) is higher for Normal text than for Knowledge text.

H2.0: The density of technical and factual information per word (i) is higher for Normal text than for Knowledge text.

H3.0: The average number of sentences per complex of sentences (V_a) is higher for Knowledge text than for Normal text.

H4.0: The number of chosen word concepts is higher for Knowledge text than for Normal text.

Materials and methods

Knowledge texts in general

In this work, we understand "knowledge text" as a specific form of the text, which contains knowledge in an explicit form. Based on our previous research (Dömeová, Houška et al., 2008), we see production rules and their advanced version, knowledge unit, respectively, as the most suitable form to represent explicit knowledge in the text. Formally, we suggested to record knowledge unit as (Dömeová, Houška et al., 2008)

$$KU = \{X, Y, Z, Q\}, \quad (1)$$

where X stands for a problem situation,

Y stands for the problem being solved in the problem situation X,

Z stands for the objective of solving the elementary problem,

Q stands for a successful solution of the elementary problem (result).

Even though there is no unique way to create sentences based on the production rules (Kendal, Creen, 2007), we can always express the knowledge unit in the following textual form (Dömeová, Houška et al., 2008): "*If we want to solve an elementary problem Y*

in the problem situation X in order to reach the objective Z, then we should apply the solution Q."

Quantitative characteristics of texts

In this part, we present the most commonly-used metrics characterizing different aspects of the texts (e.g. difficulty, communication ability, etc.) in quantitative indicators. Further on, the following parameters are used.

Complex text difficulty rate (Arya, Hiebert, Pearson, 2010)

$$T = T_s + T_p \tag{2}$$

where T_s is the syntactic difficulty rate,
T_p is the semantic difficulty rate.

Syntactic difficulty rate (Arya, Hiebert, Pearson, 2010)

$$T_s = 0.1\frac{N^2}{U \cdot V} \tag{3}$$

where N is the number of words,
U is the number of verbs,
V is the number of sentences.

Semantic difficulty rate (Hrabí, 2012)

$$T_p = 100\frac{P}{N} \cdot \frac{P_1 + 3P_2 + 2P_3 + 2P_4 + P_5}{N} \tag{4}$$

where P_1 is the number of common terms,
P_2 is the number of technical terms,
P_3 is the number of factographic terms,
P_4 is the number of figures,
P_5 is the number of recurring concepts,
P is the total number of terms in the text,
N is the total number of words in the text.

The following indicators are taken from (Hrabí 2012).

Coefficient of density of scientific and factual information per noun

$$h = 100\frac{P_2 + P_3 + P_4}{P} \tag{5}$$

Coefficient of density of scientific and factual information per word

$$i = 100\frac{P_2 + P_3 + P_4}{N} \tag{6}$$

Average number of adverbs per sentence

$$ADV_A^V = \frac{ADV}{V} \tag{7}$$

where ADV is the number of adverbs (adverbs of time, place, manner and cause),
V is the number of sentences.

Average number of adverbs per complex of sentences

$$ADV_A^S = \frac{ADV}{S} \tag{8}$$

where ADV is the number of adverbs (adverbs of time, place, manner and cause),
S is the number of complexes of sentences.

Hűbelová (2010) has used some basic formulas for describing the structure of text, e.g. average number of sentences per complex of sentences and average number of complexes of sentences per sentence could be one of them.

Average number of sentences per complex of sentences

$$V_A = \frac{V}{S} \tag{9}$$

where S is the number of complexes of sentences,
V is the number of sentences.

Average number of complexes of sentences per sentence

$$S_A = \frac{S}{V} \tag{10}$$

where S is the number of complexes of sentences,
V is the number of sentences.

Research sample and statistical methods used

In total, the research sample consists of 120 documents divided into two groups. 60 documents are written in a standard format for educational texts (normal texts), 60 documents contain the text of the same content, but rewritten into the knowledge format (knowledge texts). Normal texts are taken from educational or professional literature on agriculture waste processing (see the complete list at **http://pef.czu.cz/~houska/Agris_2014/Sample.pdf**) and represent one half of each pair. The other half of the pairs is represented with knowledge form (see above for its general form) of the texts,

which have been translated using the procedure presented in Houška and Rauchová (2013). An example of such pair follows:

Original text taken from a textbook on the industrial waste processing (see Enviregion, 2014, in Czech, translated by the authors):

„*The waste arisen from industry production differs in comparison with the one arisen from households in more properties. It differs in the composition influenced with the kind of the production. It can often contain elements, which are of the hazardous character for people as well as for the nature (toxic, explosive, flammable, etc.). That is the reason for special manipulation for such waste. Individual productions generate waste of different properties and thus there is no unique procedure for processing it. Waste from the chemical productions is often really dangerous and has to be modified before processing. Metallurgy also produces a large amount of dangerous waste. Food productions generate waste that could be transformed into a fertilizer and used in agriculture. Building industry can often recycle the waste in order to be re-used for the production of building materials or for building the houses.*“

Its knowledge form (the original text modified by the authors according to (1)) can be presented as follows:

“*If we consider the waste arisen from industry production and describe its properties, then it differs from the households one in more characteristics influenced with the source of the waste. If it contains elements denoted as hazardous for people or nature (toxic, explosive, flammable, etc.), then we should manipulate with the waste carefully. When we consider the industrial waste and describe its processing, we should bear in mind that each production generates a different kind of the waste, and thus there is no unique way of processing the waste. If dangerous waste is processed, the manipulation procedure should be described in detail in order to prevent the consequences to the environment, e.g. using the modification of the waste from chemical production aimed at the reduction of the content of the toxic metals, such as cadmium, nickel, lead, etc. When we deal with the waste processing and aim at exploiting the maximum value obtained from the waste, then we can e.g. transform the food production waste into fertilizers, building production waste into building material, etc.*”

The complete research sample (all pairs of normal and knowledge texts in Czech) is available at: **http://pef.czu.cz/~houska/Agris_2014/Sample.pdf.**

For the purposes of semantic analysis of the sample, the texts were pre-processed manually in order to allow smooth identification of the key parameters for the analysis. The notation was as follows:

- concepts (**in bold**),
- factographic terms (underlined),
- common terms (highlighted),
- figures (underlined),
- technical terms (underlined),
- verbs (underlined) and
- recurring concepts (*in italics*).

Furthermore, the texts were pre-processed for syntactic analysis, too. We distinguished:

- simple sentences (single underlined) and
- complex sentences (double underlined).

We use the indicators of descriptive statistics, such as mean, variance, standard deviation, etc. to identify basic differences among the variables presented above for normal and knowledge texts. Furthermore, we use the paired sample t-test to confirm or reject the operational hypotheses on the equivalency of individual variables for normal and knowledge texts. Using the paired version of the t-test, we respect the natural dependence among the items in both sets, where the knowledge texts were directly derived from the normal ones. See Wetcher-Hendricks (2011) for the description of these methods in details. All calculations are processed using the statistical software Statistica, version 12.

Results and discussion

First we calculate basic descriptive statistics for all partial variables, separately for normal and knowledge texts, see Table 1.

Inspired by the working hypotheses formulated in Introduction (H1 – H4) and data in Table 1, we aim at testing the following operational hypotheses.

H1.1: There is no difference in the mean value of the number of words between Normal text and Knowledge text.

H1.2: There is no difference in the mean value of the number of verbs between Normal text and Knowledge text.

	Normal text					Knowledge text				
Variable	Mean	Minimum	Maximum	Variance	Standard deviation	Mean	Minimum	Maximum	Variance	Standard deviation
N	249.8	161	311	1064.1	32.62	255.7	194	335	1098.0	33.1
U	26.1	8	41	67.9	8.24	25.6	10	43	63.4	8.0
S_s	7.5	2	12	8.5	2.91	3.7	0	11	6.3	2.5
$S_c^{(2)}$	4.6	0	9	6.8	2.61	4.1	0	9	5.4	2.3
$S_c^{(3)}$	2.2	0	5	2.2	1.49	2.7	0	7	2.8	1.7
$S_c^{(4)}$	0.5	0	2	0.6	0.77	0.8	0	4	0.9	1.0
$S_c^{(5)}$	0.2	0	1	0.2	0.40	0.4	0	2	0.3	0.5
$S_c^{(6+)}$	0.1	0	1	0.1	0.30	0.3	0	3	0.4	0.6
S_c	7.6	2	13	11.2	3.34	8.1	2	15	9.0	3.0
V_A	2.4	0.4	3.3	0.3	0.54	2.7	2	4.4	0.2	0.5
P	2.7	0	6	3.3	1.83	9.5	2	17	10.4	3.2
P_1	67.2	39	94	183.9	13.56	66.6	47	92	121.1	11.0
P_2	12.4	0	75	104.3	10.21	11.2	0	27	36.7	6.1
P_3	3.8	0	14	9.7	3.11	3.8	0	14	9.7	3.1
P_4	4.5	0	18	10.5	3.24	4.5	0	18	9.8	3.1
P_5	10.3	1	19	18.4	4.29	10.32	1	19	18.4	4.3

Note: S_s … number of simple sentences;
$S_c^{(i)}$ ..number of complex sentences consisting of i simple sentences.
Source: own processing

Table 1: Basic descriptive statistics for normal and knowledge texts.

H1.3: There is no difference in the mean value of the syntactic difficulty rate between Normal text and Knowledge text.

H1.4: There is no difference in the mean value of the semantic difficulty rate between Normal text and Knowledge text.

H2.1: There is no difference in the mean value of the coefficient of density of scientific and factual information per noun between Normal text and Knowledge text.

H2.2: There is no difference in the mean value of the coefficient of density of scientific and factual information per word between Normal text and Knowledge text.

H3.1: There is no difference in the mean of the number of simple sentences between Normal text and Knowledge text.

H3.2: There is no difference in the mean of the number of complex sentences with 2 sentences between Normal text and Knowledge text.

H3.3: There is no difference in the mean of the number of complex sentences with 3 sentences between Normal text and Knowledge text.

H3.4: There is no difference in the mean of the number of complex sentences with 4 sentences between Normal text and Knowledge text.

H3.5: There is no difference in the mean of the number of complex sentences with 5 sentences between Normal text and Knowledge text.

H3.6: There is no difference in the mean of the number of complex sentences with more than 5 sentences between Normal text and Knowledge text.

H3.7: There is no difference in the mean of the number of complex sentences in total between Normal text and Knowledge text.

H3.8: There is no difference in the mean of the average number of sentences per complex of sentences between Normal text and Knowledge text.

H4.1: There is no difference in the mean of the number of chosen words between Normal text and Knowledge text.

H4.2: There is no difference in the mean of the number of common words between Normal text and Knowledge text.

H4.3: There is no difference in the mean of the number of technical term words between Normal text and Knowledge text.

H4.4: There is no difference in the mean of the number of factographic terms between Normal text and Knowledge text.

H4.5: There is no difference in the mean of the number of figures between Normal text and Knowledge text.

H4.6: There is no difference in the mean of the number of recurring concepts between Normal text and Knowledge text.

Note: Working hypotheses and operational hypotheses do not form a hierarchy. For instance, there is no intention to understand hypotheses H4.1 – H4.6 as the particularization of the hypothesis H4.0. They are only of the same kind of the analysis (i.e. semantic difficulty of the text).

The following Table 2 shows the results of the paired t-test for dependent samples and the decision on whether we reject the above-presented null hypotheses, or not.

As indicated by Table 2, both forms of texts differ significantly in the following aspects:

N … number or words, $N(\mathrm{KT}) > N(\mathrm{NT})$;

T_S … syntactic difficulty rate, $T_S(\mathrm{KT}) < T_S(\mathrm{NT})$;

T_P … semantic difficulty rate, $T_P(\mathrm{KT}) < T_P(\mathrm{NT})$;

S_S … number of simple sentences, $S_S(\mathrm{KT}) < S_S(\mathrm{NT})$;

$S_c^{(4)}$ ..number of complex sentences containing 4 simple sentences, $S_c^{(4)}(\mathrm{KT}) > S_c^{(4)}(\mathrm{NT})$;

$S_c^{(5)}$ ..number of complex sentences containing 5 simple sentences, $S_c^{(5)}(\mathrm{KT}) > S_c^{(5)}(\mathrm{NT})$;

V_A … number of complex sentences, $V_A(\mathrm{KT}) > V_A(\mathrm{NT})$;

P … number of simple sentences per complex sentence, $P(\mathrm{KT}) > P(\mathrm{NT})$.

Variable	Type of text	Mean	Standard deviation	t-test value	P value	Hypothesis	Validity α = 0.05
N	normal	249.8000	32.62	-2.0493	0.044884	H1.1	rejected
	knowledge	255.6667	33.13				
U	normal	26.1167	8.24	0.6404	0.524364	H1.2	not rejected
	knowledge	25.5833	7.96				
T_s	normal	28.0115	25.94	4.149494	0.000108	H1.3	rejected
	knowledge	13.5023	10.10				
T_p	normal	22.0048	9.96	2.277690	0.026384	H1.4	rejected
	knowledge	19.7780	7.35				
h	normal	20.9395	10.22	1.2690	0.209409	H2.1	not rejected
	knowledge	20.2580	8.26				
i	normal	8.27834	4.88	1.7473	0.086172	H2.2	not rejected
	knowledge	7.63884	3.19				
S_s	normal	7.5167	2.90	9.7706	0.000000	H3.1	rejected
	knowledge	3.6667	2.50				
$S_c^{(2)}$	normal	4.5833	2.61	1.7810	0.080057	H3.2	not rejected
	knowledge	4.0833	2.31				
$S_c^{(3)}$	normal	2.2167	1.48	-1.9285	0.058608	H3.3	not rejected
	knowledge	2.6500	1.68				
$S_c^{(4)}$	normal	0.4833	0.77	-2.8013	0.006872	H3.4	rejected
	knowledge	0.7833	0.92				
$S_c^{(5)}$	normal	0.2000	0.40	-2.2560	0.027792	H3.5	rejected
	knowledge	0.3500	0.54				
$S_c^{(6+)}$	normal	0.1000	0.30	-1.8352	0.071522	H3.6	not rejected
	knowledge	0.2500	0.62				

Source: own processing

Table 2: Statistical analysis with the paired sample t-test.

Variable	Type of text	Mean	Standard deviation	t-test value	P value	Hypothesis	Validity α = 0.05
S_c	normal	7.5833	3.34	-1.8112	0.075194	H3.7	not rejected
	knowledge	8.1167	3.00				
V_A	normal	2.3930	0.54	-4.2878	0.000068	H3.8	rejected
	knowledge	2.7281	0.47				
P	normal	2.7000	1.82	-14.0022	0.000000	H4.1	rejected
	knowledge	9.5167	3.22				
P_1	normal	67.2167	13.55	0.6034	0.548555	H4.2	not rejected
	knowledge	66.5833	11.00				
P_2	normal	12.3833	10.21	1.3313	0.188226	H4.3	not rejected
	knowledge	11.2333	6.05				
P_3	normal	3.7966	3.11	0.3308	0.741982	H4.4	not rejected
	knowledge	3.7797	3.10				
P_4	normal	4.5167	3.24	0.0000	1.000000	H4.5	not rejected
	knowledge	4.5167	3.12				
P_5	normal	10.3167	4.29	-0.2346	0.815359	H4.6	not rejected
	knowledge	10.3167	4.29				

Source: own processing

Table 2: Statistical analysis with the paired sample t-test (continuation).

As more of the parameters shown above are correlated (e.g. if the number of complex sentences is higher for knowledge texts, we can assume that the number of words is also higher for knowledge texts, etc.), we visualize the comparison using box plots for the selected ones only (see Figure 1).

By applying the same approach to confirming the validity of the original working hypotheses H1.0 – H4.0 on the differences in characteristics among normal and knowledge texts, we obtain the results presented in Table 3.

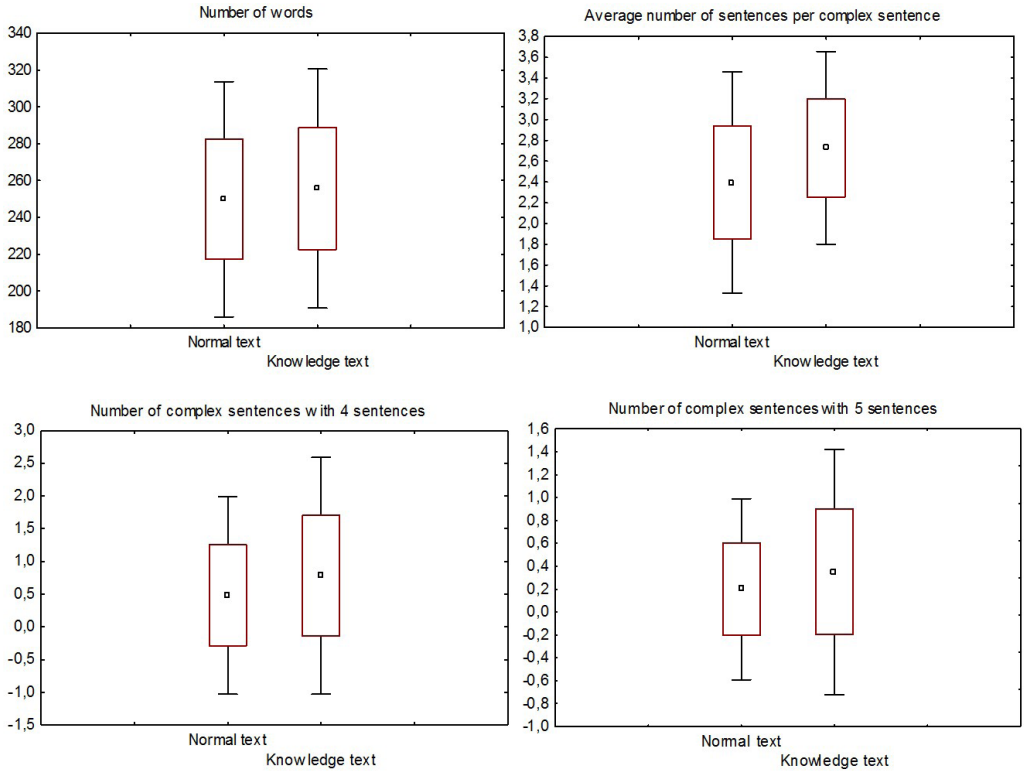
Except the H2.0 hypothesis on the differences in coefficients of density of technical and factual information between normal and knowledge texts, all working hypotheses are rejected for $\alpha = 0.05$. Our comments and the comparison with the works of other authors follow.

H1.0: There is statistically significant difference in the complex text difficulty rate between normal and knowledge texts. Normal texts achieve higher value than the knowledge ones.

At the first glance, it does not make sense. The authors, who are dealing with measuring the difficulty of texts in textbooks (e.g. McCrory, Stylianides (2014) or Miller (2011)), also show the dependence between the amount of knowledge in the text and the complex text difficulty rate as "the higher is the amount of knowledge in the text, the higher is the difficulty of the text". Explanation can be found in the way of calculating the complex text difficulty rate T as the sum of syntactic difficulty rate T_S and semantic difficulty rate T_P, see Eq. (2-4). Based on the rejected validity of the operational hypotheses H1.3 and H1.4 (both T_S and T_P values are significantly lower for knowledge texts than for normal texts), it is natural that the value of the complex text difficulty rate T is also lower for knowledge texts.

H2.0: There is no statistically significant difference in the coefficient of density of technical and factual information between normal and knowledge texts.

In contrast to our preliminary results (Rauchová et al., 2014), we have not confirmed the assumption on the differences between the texts in that characteristics. It is natural that the coefficient of density of scientific and factual information per noun h is independent on the style of the text. The number of nouns is always similar to the number of the terms in the text (see Eq. (5)). The main discrepancy between the preliminary research and the current results is caused by the coefficient of density of scientific and factual information per word i. Obviously, the variance played an important role in our preliminary research (see mean values and standard deviations for normal and knowledge texts in Table 2 for the parameter i) and roughly influenced our estimations.



Source: own processing

Figure 1: Box plots for selected parameters of normal and knowledge texts.

Variable	Type of text	Mean	Standard deviation	t-test value	P value	Hypothesis	Validity $\alpha = 0.05$
Complex text difficulty rate	normal	50.078	29.532	4.67358	0.000018	H1.0	rejected
	knowledge	33.333	13.823				
Coefficient of density of technical and factual information	normal	20.849	10.188	1.28765	0.202894	H2.0	not rejected
	knowledge	20.172	8.280				
Average number of simple sentences per complex sentence	normal	2.393	0.544	-4.2878	0.000068	H3.0	rejected
	knowledge	2.728	0.472				
Number of chosen word concepts	normal	2.700	1.825	-14.002	0.000000	H4.0	rejected
	knowledge	9.517	3.223				

Source: own processing

Table 2: Statistical analysis with the paired sample t-test (continuation).

H3.0: There is statistically significant difference in the average number of simple sentences per complex sentence between normal and knowledge texts. Knowledge texts achieve higher values than the normal ones.

This result is natural. We decompose the knowledge texts based on a formal model of the knowledge unit and its language form, respectively, see Eq. (1). The sentence always consists of two simple sentences expressing both antecedent and consequent parts of the unit at minimum. It is sometimes necessary to explain some part of the knowledge unit in more detail; as a result, the number of simple sentences becomes greater. This goes in line with mainstream literature on knowledge management or knowledge engineering. All authors, whose works we have studied, understand knowledge as enhanced data or information. This idea is really common nowadays and more applications, e.g. in agriculture (Rydval et al., 2014) or project management (Mochida, 2011) are based on it. Obviously, more words and simple sentences are required in order

to express knowledge than a statement containing information or even data only.

H4.0: There is statistically significant difference in the number of chosen word concepts between normal and knowledge texts. Knowledge texts achieve higher values than the normal ones.

In contrast to operational hypotheses H4.2 – H4.6, which concentrate on common terms, facts, technical terms, etc., the H4.0 hypothesis works with words typical for language expressions of knowledge units, mainly connectives (if, when, to, then, in order to, etc.). Here we can prove that even if there is no difference in the content of the statement (hypotheses H4.2 – H4.6 were not rejected), it can play an important role when electronic educational documents are assigned with metadata (Šimek et al., 2012), because there is no need to accompany the change of the text style with the change of the metadata. On the other hand, the formal structure of the knowledge text is too unique for the statistical analysis to be able to distinguish among normal and knowledge texts.

Conclusion

In this paper we analysed a relevant sample of educational texts on agriculture waste processing in order to investigate the differences among their normal and knowledge form. Compared to normal text, the knowledge text is characterized with sentences of more words (in average), higher occurrence of complex sentences to express the complete knowledge as well as relatively higher number of simple sentences per the complex sentence (again, in average). Particular word concepts and the intensity of their occurrence in the knowledge text allow us to differentiate both forms of text.

Several parameters, which can be used for distinguishing the texts, could serve for the purposes of further research on classification of general text as normal text or knowledge text and calculating the rate of correspondence of general text to knowledge text, respectively. In literature, we can find many kinds of analyses on document type classification (popular, narrative, scientific, etc.) or sentiment analyses of the content of documents (see e.g. Feldman, 2013 for systematic review of the current state in this area). Our results allow us to define a new type of such analysis.

Another important issue for further research is the readers' point of view. Even though we can measure and calculate that the complex difficulty of knowledge texts is significantly lower than of the normal one, we have to ensure that the readers' opinion will be in line with this theoretical assumption. Thus we are carrying out the experiment on perceiving the differences among the texts by human readers – practitioners working in agriculture and being responsible for agriculture waste processing. When these two connecting questions are answered, we will be able to evaluate the practical impacts of our theoretical findings achieved in this work.

Acknowledgements

The paper is supported by the grant project of the Internal Grant Agency of the FEM CULS Prague "Determining the quantitative characteristics of knowledge texts", No. 20131020, and by the grant project of the Internal Grant Agency of the CULS Prague "Measuring the efficiency of knowledge transfer in the agricultural waste processing sector", No. 20131001.

Corresponding author:
doc. Ing. Milan Houška, Ph.D.
Department of Systems Engineering, Faculty of Economics and Management
Czech University of Life Sciences Prague, Kamýcká 129, 165 21 Prague 6 – Suchdol
Czech Republic

Phone: +420 224 382 355, E-mail: houska@pef.czu.cz

References

[1] Arya, D., Hiebert, E., Pearson, P. The Effects of Syntactic and Lexical Complexity on the Comprehension of Elementary Science Texts. International Electronic Journal of Elementary Education, 2011, 4, No. 1, p. 107 – 125. ISSN 13079298.

[2] Asaishi, T. An Analysis of the Terminological Structure of Index Terms in Textbooks. Procedia – Social and Behavioral Sciences, 2011, 27, p. 209 – 217. ISSN 1877-0428.

[3] Cortina-Borja, M., Chappas, C. A Stylometric Analysis of Newspapers, Periodicals and News Scripts. Journal of Quantitative Linguistics. 2006, 13, No. 2 – 3, p. 285 – 312. ISSN: 0929-6174.

[4] Dömeová, L., Houška, M., Houšková Beránková, M. Systems Approach to Knowledge Modelling. Hradec Králové: GSOC. 2008. ISBN 978-80-86703-30-5.

[5] Duric, A., Song, F. Feature Selection for Sentiment Analysis Based on Content and Syntax Models. Decision Support Systems, 2012, 53, No. 4, p. 704 – 711. ISSN 0167-9236.

[6] Enviregion, Textbook on Environmental Education. 2014. [On-line]. Available: http://ucebnice3.enviregion.cz/ [Accessed: 15 May 2014].

[7] Feldman, R. Techniques and Applications for Sentiment Analysis: The Main Applications and Challenges of One of the Hottest Research Areas in Computer Science. Communications of the ACM, 2013, 56, No. 4, p. 82 – 89. ISSN 0001-0782.

[8] Graham, D. J., Hughes, J. M., Leder, H., Rockmore, D. N. Statistics, Vision, and the Analysis of Artistic Style. Wiley Interdisciplinary Reviews: Computational Statistics, 2012, 4, No. 2, p. 115 – 123. ISSN 1939-5108.

[9] Houška, M., Rauchová, T. Methodology of Creating the Knowledge Text. Proceedings of the Conference Efficiency and Responsibility in Education, 2013, p. 197 – 203. ISBN 978-80-213-2378-0.

[10] Hrabí, L. Natural Science Textbooks for the Fourth Grade and their Text Difficulty. Envigogika, 2012, 7, No. 2, p. 1 – 7. ISSN 1802-3061.

[11] Hűbelová, D. Analyses of Textbooks on Regional Geography for Primary School. Geographical Information, 2010, 14, No. 1, pp. 55 – 63. ISSN 1337-9453.

[12] Kendal, S. L., Creen, M. An Introduction to Knowledge Engineering. London: Springer. 2007. ISBN 978-1-84628-667-4.

[13] McCrory, R., Stylianides, A. J. Reasoning-and-proving in Mathematics Textbooks for Prospective Elementary Teachers. International Journal of Educational Research, 2014, 64, p. 119 – 131. ISSN 0883-0355.

[14] Miller, D. ESL Reading Textbooks vs. University Textbooks: Are We Giving Our Students the Input They May Need? Journal of English for Academic Purposes, 2011, 10, No. 1, p. 32 – 46. ISSN 1475-1585.

[15] Mochida, S. Knowledge Mining for Project Management and Execution. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2011, 15, No. 4, p. 454 – 459. ISSN 1343-0130.

[16] Newton, D. P., Newton, L. D. A Procedure for Assessing Textbook Support for Reasoned Thinking. Asia-Pacific Education Researcher, 2009, 18, No. 1, p. 109 – 115. ISSN 0119-5646.

[17] Rauchová, T., Houška, M., Luhanová, K., Černíková, K. Comparative Analysis of Quantitative Indicators of Normal and Knowledge Texts. Proceedings of the Conference Distance Learning in Applied Informatics, 2014, p. 621 – 632. ISBN 978-80-7478-497-2.

[18] Rydval, J., Bartoška, J., Brožová, H. Semantic Network in Information Processing for the Pork Market. Agris On-line Papers in Economics and Informatics, 2014, 6, No. 3, p. 59 – 67. ISSN 1804-1930.

[19] Šimek, P., Vaněk, J., Očenášek, V., Stočes, M., Vogeltanzová, T. Using Metadata Description for Agriculture and Aquaculture Papers. Agris On-line Papers in Economics and Informatics, 2012, 4, No. 4, p. 79 – 90. ISSN 1804-1930.

[20] Wetcher-Hendricks. D. Analyzing Quantitative Data: An Introduction for Social Researchers, John J. Wiley & Sons. Inc, Hoboken, New Jersey. 2011. ISBN 978-0-470-52683-5.