From Digitisation Process to Terminological Digital Resources

S. Seljan*, I. Dunđer ** and A. Gašpar***
* University of Zagreb - Faculty of Humanities and Social Sciences,
Department of Information and Communication Sciences, Zagreb, Croatia
* sanja.seljan@ffzg.hr
** ivandunder@gmail.com
*** ginasplit@yahoo.com

Abstract - Monolingual and multilingual terminology and collocation bases represent valuable additional electronic resources, which can be used in further research, in written communication and in everyday communication. Building of such resources can be supported by terminology extraction tools relying on statistical or language approaches, or on hybrid model, but require considerable human expertise in evaluation and final compilation. The paper describes the whole process: from digitisation of printed material, OCR techniques, sentence alignment and creation of translation memories, up to terminology extraction and evaluation. The performance of tools and applied methodology is assessed through standard statistical measures of precision, recall and F-measure. Experimental results are produced, deficiencies of semi-automatic statistical and linguistic system highlighted and recommendations for further research suggested.

***Keywords* – digitization, term and collocation extraction, Multi-Word Unit (MWU), statistical and language approaches, evaluation, English, Croatian**

I. INTRODUCTION

Monolinugal and multilingual terminology and collocation bases represent valuable additional electronic resources, which can be used in further research, in written communication and in everyday communication. The manual compilation of such resources is time consuming, cost-intensive and highly subjective process. Building of such resources can be supported by terminology extraction tools relying on statistical or language approach, as syntactic parsing in [12], or on hybrid model, but require human expertise in evaluation and final compilation.

Translation memory created out of sentenced aligned parallel corpus can be further used for semi-automatic extraction of domain specific terms. Hybrid approach is applied combining: statistical methods based on frequency of term candidates and differences between general and specialised corpora, and linguistic ones based on morphological and syntactic analyses.

Evaluation of extracted terminology and collocation candidates is usually based on precision and recall, but greatly depends on the purpose and usability in practice [9]. In the evaluation process, automatically extracted terms and collocations are evaluated in comparison with manually created reference list based on expert's knowledge and experience and used as „gold standard". The paper describes the whole process: from digitisation of printed material, OCR techniques and its evaluation, through sentence alignment and creation of translation memories.

The lists of automatically produced term candidates are validated by a linguist and compared to the *gold standard* made of manually annotated text. The performance of tools and applied methodology is assessed through standard statistical measures of precision, recall and F-measure. Experimental results are produced, deficiencies of semi-automatic statistical and linguistic system highlighted and recommendations for further research suggested.

II. RELATED WORK

Bases of terminology and collocations can be used independently or integrated, as in [6], [5] with visualizing environment, in machine translation (MT) systems [13], [10] or in information retrieval. Reference [8] has showed that bilingual term entries extracted from domain-specific parallel texts introduce more improvements in MT systems than specialized technical dictionary covering broader domain.

Although multi-word units are composed of two or more orthographic words (linked by dash, conjunction, or blank), they are treated as a single grammatical unit. Multi-word (MW) units can include foreign expressions (e.g., *ad hoc*), prepositions (e.g. *freeze up*, *depend on*), adverbs (e.g., *of course*), idiomatic noun constructions (e.g., *know how*, *per cent*), expressions (e.g., *well being*), as on BNC (British National Corpus) web-page.

Multi-word terminology directly express the concept, as in Sager's lit of requirements [2], does not overlap with other terms, it is lexically systematic, does not contain unnecessary information, and is independent from the context. Terms are congruent to general rules of word-formation. Multi-word terms are composed of two or more words or compound words that are concept-oriented and used in a specific domain.

Collocations are defined a sequence of words or terms that appear more often than would be expected by chance

(e.g. powerful computers/ *strong computers; fast train/ *quick train; fast food/ *quick food; quick meal/ *fast meal; strong tea/ *powerful tea), as in [3]. Collocations represent a high domain of interest in machine translation and considerable effort has been made to perform suitable techniques for collocation extraction from parallel corpora [6], [11]. *Idioms*, i.e. combination of words having figurative meaning, are not included in this base.

As in [3], collocations are characterized by noncompositionality (meaning of the collocation can not be predicted from the meaning of the parts), by non-substitutability (components can not be substituted), and by non-modifiability (not modified through additional lexical material of grammatical transformations). According to [13], collocations are considered to be a subset of *multi-word expressions* that constitute arbitrary conventional associations of words within a particular syntactic configuration.

III. RESEARCH

A. Data Set

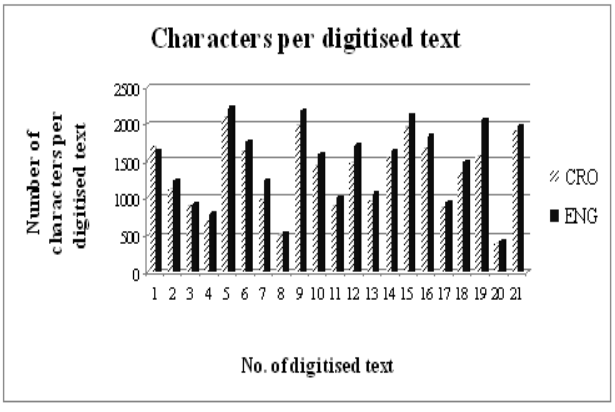
The research is performed on parallel bilingual Croatian-English abstracts relating to the philosophical and sociological topics of society, religion, dignity, freedom, peace, etc. The process includes: digitisation of printed material, OCR techniques and its evaluation; sentence alignment and creation of translation memories; extraction of terminology and collocation using different types of tools based on statistical and language approaches; evaluation and analysis.

This research is made on 21 parallel Croatian-English texts from the philosophical-sociological-religious domain. On average, English texts are composed of 11,97% more characters.

TABLE I. DATA SET

	#characters	***#words***	***#abstracts***	***Arith. mean #characters***
Cro	27 287	3 965	21	1 299,38
Eng	30 354	4 881	21	1 445,43

CHART I. CHARACTERS PER DIGITISED TEXT



B. Methods and Tools

Digitisation and data preparation

For conducting this research, data in form of Croatian and English texts had to be digitised with a scanner. Digitisation means recording, storing and processing content using digital cameras, scanners and computers [1]. It is the process of creating a digital representation of an object, image, document or a signal, enabling them to be stored, displayed, disseminated and manipulated on a computer.

As scanners generate a raster image or a snapshot of an object, Optical character recognition (OCR) software was applied in order to extract, edit, search and repurpose data from scanned objects.

OCR software recognizes text by analysing the structure of the object that needs to be digitised, by dividing it into structural elements and by distinguishing characters through comparison with a set of pattern images stored in a database and built-in dictionaries.

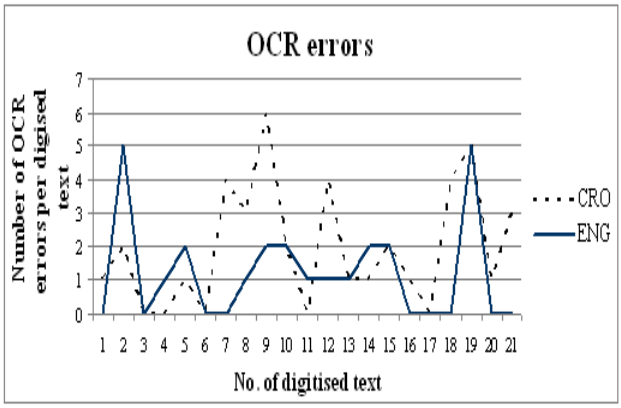
For the purpose of digitisation a HP Scanjet G3110 flatbed scanner was used, set to 300 dpi, grayscale scanning and other default settings. Scanned documents were in A5 format and text was written in Times New Roman font, size 10, standard black font colour on white background.

Optical character recognition was carried out by using Abby Fine Reader 8.0.0.677, which allowed conversion of scanned input texts from bitmap format to encoded text. Errors are unavoidable in advanced computer vision applications such as Optical character recognition, and the noise induced by these errors presents a serious challenge to downstream processes that attempt to make use of such data [7].

Typical errors that occurred during Optical character recognition were misrecognitions of characters, e.g. (,) → (.), (i) → (;), missing whitespace characters ( ) or apostrophes ('), various forms of substitution errors, e.g. (l) → (i), (h) → (ll), (rn) → (m), («)→ ((() as well as space deletion, e.g. (*među crkvenim*) → (*međucrkvenim*) and insertion errors, e.g. (*neučinkovitost*) → (*ne učinkovitost*).

The most frequent OCR errors in Croatian were substitution errors (e.g. (l) → (i)) and space deletion, where two words were unified in one by mistake. The most frequent OCR errors (83%) in Croatian were substitution errors (e.g. (l) → (i)) and space deletion, where two words were unified in one by mistake ((*među crkvenim*) → (*međucrkvenim*)). The most frequent OCR errors (87%) in English were substitution errors (e.g. («)→ ((()) and missing apostrophes (').

CHART II. OCR ERRORS



Arithmetic mean of OCR errors in digitised Croatian texts is 1,95 and in digitised English texts 1,19. OCR errors have an impact on later-stage processing and usability, so all scanned texts were post-edited afterwards.

Extraction Tools

Multi Term Extract - SDL Trados

The tool used for terminology extraction MultiTerm Extract offers a variety of extraction possibilities (set up a statistical threshold, min and max term length, min translation frequency, and max number of translations). It can be used for the extraction from monolingual or bilingual translation memories. For each term candidate the program suggests a number of probable translations presented in a term candidate list on a user-friendly graphic interface.

During the term extraction process it uses stop-list containing functional words, such as articles, conjunctions, prepositions, etc. in order to refine the suggested list of term candidates. After validating terms and their translations it is possible to export them to MultiTerm XML or a tab delimited format. This program is interrelated with other programs in SDL Trados package designed to assist to the translators before and during translation, but can be also used to create independent terminology list.

SDL PhraseFinder

SDL's PhraseFinder uses sophisticated algorithms to understand the structure of a language and identify candidate terms.

Term extraction is based on linguistical and statistical approach and does not support Croatian language. It analyzes the content of monolingual and bilingual files identifying single-word and multiple-word terminology candidates which can be viewed, edited and validated in the PhraseFinder screen. It supports various file formats (HTML, HTM, RTF, SDLX Translation Document (ITD file), SDLX Translation Memory (MDB file), TXT) and languages (English, French, German etc.).

The extracted terminology candidates can be reviewed to identify valid terms and ordered by frequency, rank or alphabetically. Validated terms can be saved in (.phr) format and exported in TXT format for use in other applications. Filter and display settings refer to the following parameters: term length, number of context sentences, ignoring of function words, numbers, first or all uppercases, unfound text, duplicate capitalisation and sort of term candidates.

NooJ

Term extraction was performed by a linguistically-based environment NooJ developed by Silberztein in 2004 at the University Franche-Comté Paris, France. NooJ's multilingual engineering platform provides specific tools for the formalization of linguistic phenomena at the level of orthography, morphology, syntax, semantics, and lexicon. It can parse and process large texts and corpora, lounge sophisticated queries in order to produce various results (concordances, statistical analyses, information extraction, etc.).

A set of local regular grammars is required to perform multi-word units extraction. Based on large-coverage dictionaries and grammars, NooJ linguistic engine can process texts and corpora in real time regardless of file formats (varying from MS Office, HTML and PDF to XML documents).

IV. EVALUATION LISTS

Evaluation of automatically or semi-automatic created lists is performed on the basis of comparison with regard to human-created term list based on human knowledge, experience and intuition. Results are analyzed using standard measures of recall, precision and F-measure.

Reference list

For the purpose of this research a reference list is created in order to assess tools' performance. It is made on the basis of 21 abstracts, out of 42 in total.

The human reference lists consist of 491 terms in English and 506 terms in Croatian. These reference lists contain not only terminology specific for the subdomain, but also terms that more frequently appear but belong to everyday language. The lists contain mainly nominal multi-word units and very rarely collocations.

Structuring the reference list is a demanding task because some expressions in Croatian language are paraphrased in English or translated by anaphoric expressions.

Although, the focus of expert's extraction is to detect bilingual phrase pairs which meet linguistic and terminological criteria and domain-specific coverage, some terms in English language were rejected because of a discrepancy caused by Croatian syntax.

Disregarded terms are for example (*pripadnici društva / members of its society; doseg svjetovnih odgovora / scope of their worldly responses; poprište se njihova sukoba / focus of conflict; svijet i za društvo / world and society; počinjena ili proživljena zla / evil that has been done or experienced; sjećanje na pretrpljenu patnju / memory of suffering that has been experienced*).

The reference list for Croatian language, was supplemented by 89 false positives, i.e. terms not found in the reference list, and consists of 506 terms in total.

The reference list for English language contains 491 terms, including 95 false positives. Table 2 presents number of n-grams and their distribution in the reference list for both languages. In both languages there is the highest number of 2-grams and 3-grams, whiled in English there is higher number of 4-grams and 5-grams.

TABLE II. N-GRAMS OF THE REFERENCE LIST

N-grams total	2-grams	3-grams	4-grams	5-grams and more	# terms & collocations
Cro	254	168	68	16	**506**
Eng	166	167	91	67	**491**

The frequency of terms decreases with their length in both languages. There is bigger proportion of n-grams in Croatian language comparing to English, except for 4-, and 5 grams. Noun phrases prevail in both languages.

NooJ list

Linguistic approach is based on creation of 22 local grammars for Croatian term candidate extraction spanning up to 5-grams, where N stands for noun, A-adjective, S-preposition and C-conjunction, and on 29 local grammars for English term extraction including 7-grams, where POS are as follows: N-noun, A-adjective, PREP-preposition, DET-article and CONJ-conjunction. POS tagging is automatically performed by using NooJ tool for language engineering.

Table 3 presents local grammars for Croatian and English language.

TABLE III. N-GRAMS OF THE REFERENCE LIST

N-grams total (22)	**Local grammars for Croatian language**		
2-gr.	<N><N>;	48	Most common
	<A><N>;	208	
3-gr.	<A><N><N>;	29	
	<A><A><N>;	29	
	<N><S><N>;	12	
	<N><C><N>;	36	
4-gr.	<A><C><A><N>;	17	
3-5-gr.	<N><N><N>; <N><N><A><N>; <A><N><C><A><N>; <N><A><C><A><N>;	Least common	
N-grams total (29)	**Local grammars for English language**		
2-gr.	<N><N>;	10	
	<A><N>;	178	
3-gr	A><A><N>;	21	Most common
	<N><PREP><N>;	97	
4-gr.	<N><PREP><DET><N>;	26	
	<A><N><PREP><N>;	28	
	<N><PREP><A><N>;	23	
	<A><CONJ><A><N>;	14	
5-gr.	<N><PREP><DET ><A><N>;	15	
	<A><N><PREP><DET><N>;	15	
6-gr.	<N><PREP><DET><N><PREP><N>;	5	
3-7-gr.	<N><A><N>;	Least common	
	<N><A><PREP><N>;		
	<N><PREP><DET><N><N;		
	<N><PREP><N><PREP><DET><N>;		
	<N><A><PREP><N><DET><PREP><N>;		

As presented by table 3, 2 and 3-grams are almost equally represented term candidates in Croatian and English. While 2-grams follow the same language pattern of noun phrases possibly proceeded by adjective, 3-grams differ in language patterns. 4-5grams are significantly more represented in English, due to analytic language structure.

Local grammars are created on the basis of the most frequent POS patterns in the reference list. Local grammars differ in number because of the usage of articles and prepositions in English and inflection in Croatian language. Also, a small, high-level priority dictionary for English language is created in order to avoid ambiguities. For example the word *'a'* which always refers to the determiner not the noun (in abstracts) is stored in high-priority dictionary together with other words which correspond to more than one lexical entry, as presented in table 4.

TABLE IV. HIGH- PRIORITY DICTIONARY FOR ENGLISH LANGUAGE

a, DET	for, PREP
an, DET	on, PREP
the, DET	under, PREP
while, CONJ	down, PREP
if, CONJ	above, PREP
in, PREP	after, PREP

Lexical constraint is set for the noun to be in the nominative case at the beginning of the phrase (<N+Case=Nom-Type=Kr>).

V. RESULTS

Statistical analysis is based on statistical measures of *precision* (the proportion between valid automatically extracted terms and all automatically extracted terms), *recall* (the proportion between valid automatically extracted terms and manually extracted terms), and the *F measure* (the ratio between *precision* and *recall*). The

results of automatic term extraction by NooJ are presented in Table 5.

TABLE V. THE RESULTS OF AUTOMATIC TERM EXTRACTION

	NooJ		MultiTerm Extract		PhraseFinder	
	Cro	Eng	Cro	Eng	Cro	Eng
No. of terms	1004	1117	118	118	-	325
Validated terms	316	240	42	42	-	122
Precision (%)	31%	21%	**36%**	**36%**	-	**38%**
Recall (%)	**62%**	**49%**	8%	9%	-	**25%**
F1 (%)	**42%**	30%	13%	14%	-	**30%**

Table 6 presents the most and the least frequent POS-patterns of the reference list and list of validated Croatian and English terms obtained automatically using local grammars. POS patterns which frequency is lower or equal to 5% are: NPAN, ANPDetN, NPDetNPN, NPDetANPN.

TABLE VI. POS-PATTERN STATISTICS OF THE REFERENCE LIST AND LIST CREATED BY NOOJ

Reference list			**List created by NooJ**	
	Cro	Eng	Cro	Eng
AN	186 (37%)	139 (28%)	208 (42%)	178 (37%)
NN	47 (13%)	25 (5.0%)	48 (9.8%)	10 (2.1%)
NPN	83 (17%)	76 (15.4%)	12 (2.4%)	97 (20.6%)
NConjN	44 (8.6%)	43 (9%)	36 (7.3)	-
AAN	27 (5.3%)	22 (4.4%)	29 (5.9%)	21 (4.4%)
ANN	17 (3.3)	2 (0.4%)	29 (5.9%)	4 (0.8%)
AConjAN	2 (0.3%)	3 (0.6%)	17 (3.4%)	14 (2.9%)
ANPN	6 (1.1%)	7 (1.4%)	-	28 (5.9%)

VI. DISCUSSION

Human reference list consists namely of <AN> pattern in Croatian and English followed by <NPN> and <NN> pattern in Croatian and <NPN> and <NConjN> in English. NooJ offer more than 1000 term candidates consisting of indicated local grammars according to which all term candidates consisting of e.g. <AN> patterns for English or Croatian were extracted regardless statistical appearance or relevance. Therefore, it offered high number of candidates which were retrieved by indicated local grammars.

In regard to the number of discarded term candidates NooJ offers better for Croatian language. Language tool NooJ has significantly higher recall comparing to statistically-based tools, denoting that number of suggested term candidates overlap more with reference lists in both languages, as they mostly consist of similar language patterns.

Rule-based candidates highly depend on their identification in the text retrieved by appropriate local grammar. When comparing language-based and statistical-based tools for terminology extraction, language tool offers significantly higher number of candidates mostly composed as noun phrases: <AN> pattern in Croatian 42% and in English 37%, followed by <NN> pattern 9,8% in Croatian and in English by <NPN> pattern (20,6%). Statistically-based extraction tools would offer much better results in bigger domain specific corpora.

Verb-XX collocations represent challenge for statistically-based tools, but since the abstracts belong to various domains of philosophy, sociology and religion and are not voluminous, the collocations are not significantly retrieved as candidates.

As MultiTerm Extract tool is based on statistical frequency, it offered relatively good results consisting of terms which would be suitable for the general language, but not for this domain consisting of several subdomains.

The abstracts collected for this research belong to various subdomains and therefore the characteristic terms do not appear frequently, and are not retrieved by statistical term extraction tool. Term extraction process offered equal results when extracting from two monolingual corpora (Croatian and English) and when extracting from the bilingual translation memory.

The ratio between the extracted term candidates is 12,5:1 indicating that use stop word lists improves significantly the results. Stop word lists eliminates term candidates beginning with commas, prepositions, conjunctions and other functional words appearing at the beginning of tem candidates.

PhraseFinder tool, which is based on hybrid model offered higher score for precision and relatively high value for recall. It offers possibility for term extraction only for English, and not for Croatian language.

VII. CONCLUSION

As this collection of 21 abstract is not voluminous, statistical extraction offered less good results, especially MultiTerm Extract tool which is only statistically-based. Language-based tool NooJ gave the best results, while PhraseFinder which is based on hybrid model gave medium results for English list. PhraseFinder does not offer the possibility for Croatian terminology extraction.

For better results, more extensive corpus in the specific domain should be compiled, in order to achieve significantly better results.

The final compilation of terminology and collocation bases still highly depends on human expertise. Presented tools can serve as assisting tool, but are useful only when using voluminous texts from the specific domain and when reference list is created as intersection of several experts, with predefined purpose of use.

REFERENCES

[1] Lopresti, D. Optical Character Recognition Errors and Their Effects on Natural Language Processing, International Journal on Document Analysis and Recognition, vol. 12, no. 3, 2009, pp. 141-151.

[2] Love, S. Benchmarking the performance of Two Automated Term-extraction systems: LOGOS and ATAO. Mémoire de maîtrise, Univ. de Montréal, 2000.

[3] Manning, Chiristopher. D.; Schütze, H. Foundations of Statistical Natural Language Processing. MIT, 2002.

[4] Seljan, Sanja; Gašpar, Angelina. First Steps in Term and Collocation Extraction from English-Croatian Corpus // Proceedings of 8th International Conference on Terminology and Artificial Intelligence (TIA2009). Toulouse, France, 2009.

[5] Seretan, V; Nerima, l; Wehrli, E. A tool for multi-word collocation extraction and visualization in multilingual corpora. In Proceedings of the Eleventh EURALEX International Congress (EURALEX), 2004, p. 755–766.

[6] Seretan, Violeta. An integrated environment for extracting and translating collocations. Proceedings of 5th Corpus Linguistics Conference, Liverpool, UK, 2009.

[7] Smolčić, J.; Valešić, A. Legal Contexts of Digitization and Preservation of Written Heritage, INFuture2009 – Digital Resources and Knowledge Sharing, Zagreb, pp. 87-94, ISBN: 978-953-175-355-5

[8] Tatsuya, Izuha. Machine Translation Using Bilingual Term Entries Extracted from Parallel Texts. Systems and Computers in Japan, 36(8), 2005, pp. 23-30.

[9] Thurmair, G. Making Term Extraction Tools Usable. Joint Conference of 8th European Association for Machine Translation (EAMT) and the 4th Controlled Language Applications (CLA) Workshop, EAMT- CLAW, 2003.

[10] Turcato, Davide. Automatically Creating Bilingual Lexicons for Machine Translation from Bilingual Text. Proceedings of COLING, 1998. Bilingual Lexicons for Machine

[11] Wanner, Leo; Bohnet, Bernd; Giereth, Mark; Vidal, Vanesa. The first steps towards the automatic compilation of specialized collocation dictionaries. Application-Driven Terminology Engineering: Special issue of Terminology 11:1, 2005, pp. 143–180.

[12] Wehrli, Eric; Seretan, Violeta; Nerima, Luka. Sentence analysis and collocation identification. Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE), Beijing, China, 2010, pp. 27–35.

[13] Wehrli, E.; Seretan, V.; Nerima, L.; Russo, L. Collocations in a Rule-Based MT System: A Case Study Evaluation of Their Translation Adequacy, EMT, 2009.

[14] L'Homme, M—C. Hee S.B. A Methodology for Developing Multilingual Resources for Terminology. Proceeding of LREC - Language Resources and Evaluation, 2006.

[15] Drouin, P. Term extraction using non-technical corpora as a point of leverage. In Terminology, 2003, vol. 9, no 1, p. 99-117

[16] Barlow, M.: Collocate 1.0: Locating collocations and terminology. TX:Athelstan, 2004.

[17] L'Homme, Marie-Claude and Hee Sook Bae. A Methodology for Developing Multilingual Resources for Terminology. In Proceeding of LREC 2006. Language Resources and Evaluation, 2006.