Media monitoring and information extraction for the highly inflected agglutinative language Hungarian

Júlia Pajzs[1], Ralf Steinberger[2], Maud Ehrmann[3], Mohamed Ebrahim[4], Leonida Della Rocca[2], Eszter Simon[1], Stefano Bucci[2], Tamás Váradi[1]

(1) Hungarian Academy of Sciences, Budapest, Hungary
(2) European Commission – Joint Research Centre, Ispra, Italy
(3) Sapienza University, Rome, Italy
(4) Cognizant SetCon, Munich, Germany
pajzs.julia@nytud.mta.hu, Ralf.Steinberger@jrc.ec.europa.eu

Abstract

The *Europe Media Monitor* (EMM) is a fully-automatic system that analyses written online news by gathering articles in over 70 languages and by applying text analysis software for currently 21 languages, without using linguistic tools such as parsers, part-of-speech taggers or morphological analysers. In this paper, we describe the effort of adding to EMM Hungarian text mining tools for news gathering; document categorisation; named entity recognition and classification for persons, organisations and locations; name lemmatisation; quotation recognition; and cross-lingual linking of related news clusters. The major challenge of dealing with the Hungarian language is its high degree of inflection and agglutination. We present several experiments where we apply linguistically light-weight methods to deal with inflection and we propose a method to overcome the challenges. We also present detailed frequency lists of Hungarian person and location name suffixes, as found in real-life news texts. This empirical data can be used to draw further conclusions and to improve existing Named Entity Recognition software. Within EMM, the solutions described here will also be applied to other morphologically complex languages such as those of the Slavic language family. The media monitoring and analysis system EMM is freely accessible online via the web page http://emm.newsbrief.eu/overview.html.

Keywords: Media Monitoring; Hungarian language; information extraction and inflection.

1. Introduction

Hungarian is one of the 24 official European Union (EU) languages, spoken by roughly fifteen million speakers world-wide. Like other Finno-Ugric languages, Hungarian is morphologically extremely complex, which makes text processing and information retrieval highly challenging. Hungarian uses up to 18 cases (depending on definition), 2 numbers and several other inflectional and derivational suffixes. Hungarian uses vowel harmony, meaning that most suffixes have several forms, and the actually used form of the suffix depends on the quality of the vowel in the stem of the word. Additionally, Hungarian is an agglutinative language, meaning that morphemes can be added productively to both common and proper nouns. An example is the agglutinated name form *Obamázásaitokat*, which is the accusative of the plural genitive of the noun-to-verb plus noun-to-noun derivative form of the name *Obama* (*obamázás-a-i-tok-at* – Obama.doing-POSS-PL-2PL-ACC), as used in the sentence:

Hu: Unom már az állandó Obamázásaitokat.
En: I am fed up with your always mentioning Obama.

where 'Obama.doing' is when somebody behaves similar to Obama or speaks a lot about Obama. As a consequence of this productive language feature, there can be thousands of potential suffixed forms for the same Hungarian lemma, a fact existing Named Entity Recognition (NER) software has to consider (e.g. Varga and Simon, 2007; Szarvas et al., 2006b). The majority of Hungarian language technology systems therefore make use of lemmatisers or of morphological analyser software: either *Humor* (Novák, 2003) or the open source *HunMorph* (Trón et al. 2005). However, integrating such software in the highly multilingual EMM platform (Steinberger 2013) is not a viable option, both for licensing reasons and because integrating and maintaining third-party tools is tricky when each tool has different specifications, a different level of maturity and different update cycles.

We first present EMM's multilingual text mining components to explain the working environment for which solutions are sought (Section 2). Section 3 summarises related work in the field of Hungarian NER and name lemmatisation. We then describe the specific steps that have been carried out to adapt EMM's text analysis tool set to the Hungarian language (Section 4). This includes presenting results for the task of quotation recognition (4.2), various experiments aimed at overcoming the challenges due to the inflectional features of Hungarian (4.3), and providing detailed frequency lists of person and location name suffixes as found in real-life news text (4.4). Section 5 summarises and concludes the paper.

2. Text mining in EMM

EMM (Steinberger et al. 2009; Steinberger 2013) is a fully-automatic media monitoring platform developed entirely at the European Commission's *Joint Research Centre* (JRC). EMM gathers an average of almost 200,000

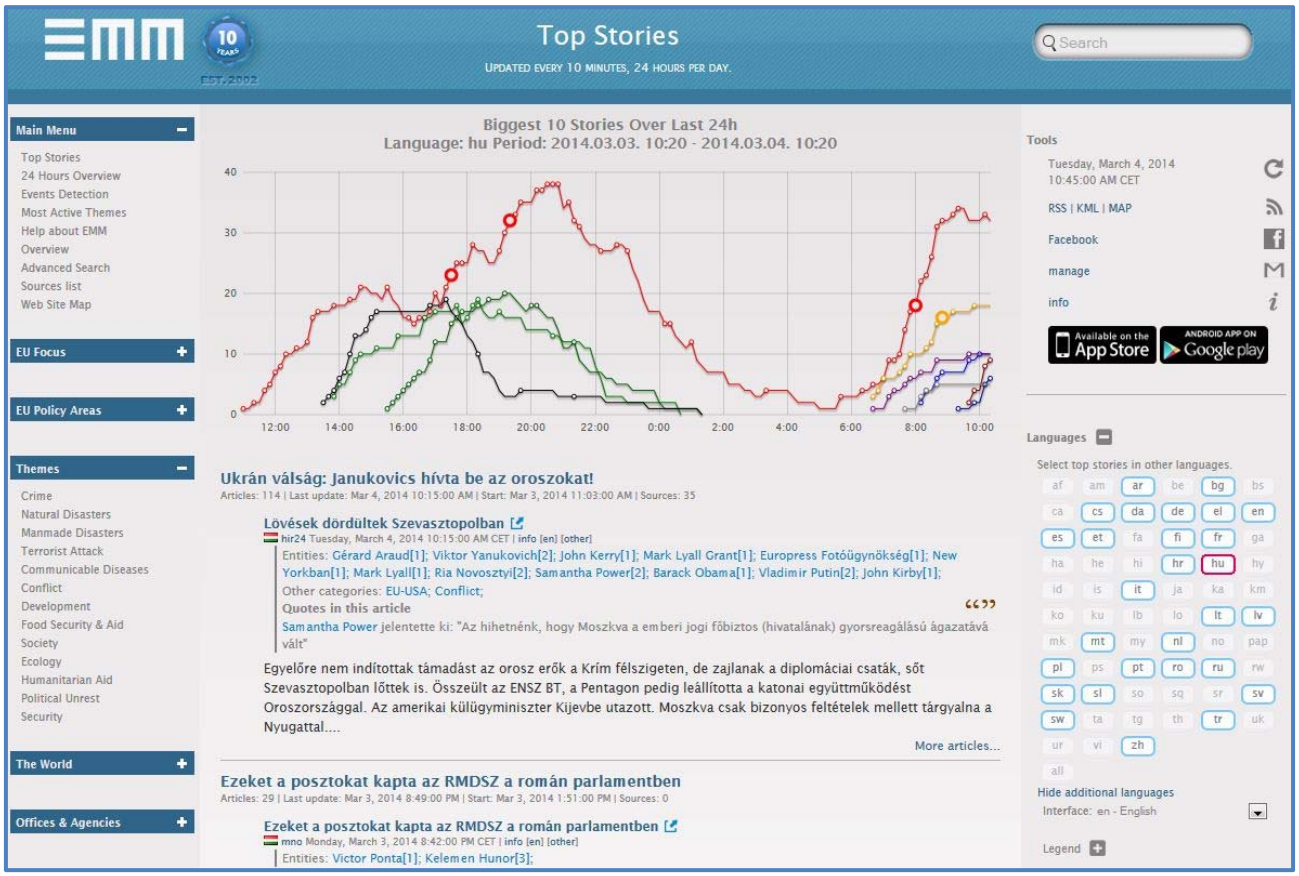

Figure 1. EMM-NewsBrief page showing the major current Hungarian language live news and their development over the last 24 hours, together with information automatically extracted from each news cluster, including news categories, names and quotations.

online news items per day from about 4,000 selected news sources around the world, by visiting these sites up to every ten minutes and extracting the news text. For text classification, EMM allows the use of Boolean search word combinations, combined with positive and negative weights and thresholds. Wild cards representing exactly one letter (_) or zero, one or more letters (%) help deal with language variants and with inflection. Related news of the same language are identified using a bottom-up hierarchical clustering process. The graph in **Figure 1** shows the ten largest news clusters and the development of their cluster size over the last 24 hours. The web page gets updated every ten minutes, allowing users to always see the latest state of news affairs. Related news across languages are identified by calculating a pair-wise cluster similarity for all currently 210 language pairs, based mostly on the overlap of multilingual subject domains and on mentions of entities (locations; persons and organisations). **Figure 2** shows the largest news clusters of a given calendar day, together with automatically generated links to the equivalent news in other languages. For this cross-lingual linking to work, it is crucial that multilingual entity variants are identified and represented by the same unique identifier. Locations in EMM are thus grounded to their geographical co-ordinates. Persons and organisations are represented in EMM by their unique numerical identifiers. Spelling variants of these names are identified by using transliteration, normalisation and string distance calculations (Pouliquen and Steinberger 2009) in order to be assigned to the same name identifier as the main name spelling. NER is performed by first looking up known names (about half a million known place names and approximately the same amount of previously recognised person and organisation names) and by then applying almost language-independent hand-written rules to recognise previously unseen person and organisation names. These NER rules make use of targeted lists of language-specific dictionaries (e.g. including lists of stop words and lists of frequent modifiers, lists of titles, professions, etc.). NER in EMM only recognises person names consisting of at least two name parts in order to allow disambiguating and grounding the names to real-life entities. The focus is not on recognising every single name mention, but to achieve high recognition precision and to achieve a good recall *at document level* because the aim is to tell the users which names are mentioned inside the news article. Quotations (reported speech) are recognised in EMM if they are mentioned adjacent to the name of the speaker and a quotation verb (e.g. *said*) (Pouliquen et al. 2007).

In order to add a new language to EMM's tool set, it is not usually necessary to add any new rules as most information extraction rules in EMM are language-independent. The language adaptation effort for EMM's information extraction tools is limited to adding the language-specific targeted dictionaries for that

Figure 2. Hungarian page of EMM-NewsExplorer, showing the biggest news clusters of the day and the countries and people mentioned that day in Hungarian news. Hyperlinks allow drilling down and viewing details. The language codes next to the cluster titles are direct links to the equivalent news in other languages. The calendar allows exploring the news of the past.

language (e.g. titles, frequent first names, reporting verbs, various types of stop words, etc.). For highly inflected languages, producing such targeted dictionaries is particularly tedious. For the classification in EMM to work, search word combinations need to be formulated. Finally, all functionality needs to be tested and tuned. For details regarding the adaptation effort to the Hungarian language, see Section 4.1. EMM does not make use of any third-party tools, of parsers, of part-of-speech taggers or of morphological analysers.

3. Related work

To our knowledge, besides EMM, there are no other fully-automatic Hungarian news analysis systems and no previous work exists on the recognition of reported speech quotations for Hungarian. We will thus focus in this section on NER, on name lemmatisation, and on methods to identify inflected variants of uninflected names from a name list.

NER requires linguistic knowledge about the structure or composition of each type of name. For example, person names usually consist of first names and last names, with optional name prefixes and suffixes, and many organisation names contain acronyms such as *Corp.* or *Ltd.* Following McDonald's (1996) terminology, elements such as *Corp.* that are part of the name are called *internal evidence*. However, there are many names that do not provide the structural indication of their category membership. Thus, to recognise and classify names, knowledge about how names appear in free text is also required. This knowledge consists of contextual clues about how each type of name may appear. For example, person names may have professional titles or descriptions preceding or following the name. These are examples of *external evidence*. Within EMM, we refer to these words as *trigger words*. Entities cannot be categorised based on internal evidence alone and require external evidence from the context as well. NER applications are thus mostly based on manually formulated or automatically learnt patterns that describe the internal structure of names, as well as context-sensitive rules which give clues for their recognition and their classification.

As for **Hungarian NER**, we have knowledge of only one rule-based NER system that has been formally evaluated (Gábor et al. 2003). Similarly to other rule-based NER systems, it makes use of both internal and external evidence. This system has been reported to achieve 82.13% overall F-measure in recognising person, location and organisation names on a 20,000 token sub-corpus of the *Szeged NER corpus* (Szarvas et al. 2006a).

We are additionally aware of two Hungarian NER systems based on machine learning methods: Szarvas et al. (2006b) published results on their NER system based on C4.5 decision trees with boosting. Their system achieved state-of-the-art performance for English, and it reached 94.77% CoNLL F-measure for Hungarian. The second NER system is called *Hunner* (Varga and Simon 2007). It is based on the maximum entropy method and reached 95.06% F-measure on the same test corpus. Note that these results refer to the mere recognition of names, which differs from the task in EMM. In EMM, (a) morphological and other name variants need to be matched to the same base form, and (b) location and person names need to be grounded to a real-world entity. For instance, the fifteen places world-wide with the identical name of *Paris* need to be distinguished, in addition to the standard task of distinguishing location names from person names (e.g. *Paris Hilton*). See Pouliquen et al. (2006) for details on the

language-independent disambiguation procedure deployed in EMM.

Statistical NER systems also use some **morphological information** (lemma, POS tags, etc.), but experiments show that morphological features do not have significant effect on the recognition accuracy: Removing them from the Hunner system decreased the F-measure from 95.06% to 94.70% (Varga and Simon, 2007). However, if the task is to ground a newly found inflected name variant to an uninflected name from a list of existing names (*list lookup*, e.g. from gazetteers of place names or from lists of known person names), as is the case in EMM, some sort of name variant mapping is required in order to match the inflected form from the text with the uninflected form in the pre-existing name list. Once matched, any knowledge contained in the list can be exploited, such as the type of the name (e.g. person vs. organisation), the geographical co-ordinates of place names, etc.

Most **lookup approaches** implicitly require candidate words to exactly match an element of a list. However, one may want to allow some flexibility in the match conditions. There are several lookup strategies used in NER. First, words can be lemmatised and only lemmas matched to list elements. For this purpose, the guesser functionality of a morphological analyser can be used, but they typically work with a limited lexicon and contain only frequent names and their lemmatisation rules. More sophisticated methods can also be applied, e.g. web-search-based heuristics for NE lemmatisation (Farkas et al. 2008), or adaptation of a morphological analyser for handling NEs (Simon 2013). However, even lemmatisers are not a solution to all inflection-related problems: Piskorski et al. (2009) tested a number of different lemmatisers on Polish named entities (as opposed to on common words) and the authors found that their accuracy on names was only between 35% and 75%. Some of the reasons they identified are that proper names partially follow different inflection patterns and that patterns for foreign names can be different from those of local names.

Second, words can be fuzzy-matched against the list using some kind of metric that measures string distance. This allows capturing small lexical variations in words that are not necessarily inflectional. For example, Tsuruoka and Tsujii (2003) calculate the edit distance between spelling variations of protein names in biomedical texts, while Cohen and Sarawagi (2004) use the Jaro-Winkler distance metric to correct mismatches. Piskorski and Sydow (2007) tested and compared a whole range of string similarity metrics. Note that these fuzzy-matching methods may help identify that various inflection forms (and other variants) belong to the same entity, but they cannot be used to identify the base form of the name, which is a requirement in EMM.

To summarise: EMM seems to be the first fully automatic system that automatically analyses Hungarian news and that extracts Hungarian reported speech quotations. There are Hungarian NER systems that perform very well. These usually require the usage of lemmatisers, but for name recognition purposes it is possible to do without. To match the inflected name forms found in text to existing lists of disambiguated named entities remains a challenge even when using lemmatisers. The solutions proposed in EMM are described in the next section.

4. Adapting EMM to the Hungarian language

In this section, we will provide details about the actual effort that was required to add Hungarian to EMM. Previous published efforts to adapt EMM to Swahili (Steinberger et al. 2011) and Arabic (Zaghouani et al. 2010) were rather different in nature. Swahili is relatively straight-forward from an Information Extraction point of view (the markers of the semantic noun classes are easy to deal with for classification and recognition purposes). The main challenges for Arabic are the non-existence of upper and lower case (which is an important feature for NER), as well as the usage of prefixes in addition to suffixes. We perceive the main challenge for Hungarian to be the complex morphology. A feature worth pointing out is the inverted order of name parts for Hungarian names in Hungarian text (surnames before first names, e.g. *Orbán Viktor*) while foreign names in Hungarian text follow the inverted order (first name before surname). This latter is the order we observe in the news in all other languages covered by EMM (e.g. *Angela Merkel*).

We will first give a concrete idea about the individual tasks that had to be performed when adding the Hungarian language to EMM (Section 4.1). We then present the evaluation results for the quotation recognition task (4.2). In Section 4.3, we present our experiments to recognise and to lemmatise Hungarian names, as well as an evaluation on the results achieved. After the named entity recognition software had been running for a few months, we produced statistics on the frequency of different name suffixes, separately for person and for location names. These will be presented in Section 4.4.

4.1 Effort to adapt EMM to Hungarian

The Hungarian version of EMM was developed as part of a collaboration agreement between the Hungarian Academy of Sciences and the JRC. The Hungarian partners spent about three months on this collaboration, which includes adding further Hungarian language news sources, writing category definitions, producing various targeted dictionaries, providing lists of the most frequent Hungarian suffixes, contributing with their know-how on Hungarian morphology and word order, evaluating the system results and suggesting improvements (see also Section 2). In addition to these three months, an estimated three months of the JRC's developers and computational linguists were spent. These six months are about double the usual effort of adding a new language to EMM. The reasons are that Hungarian is more complex and that several experiments were carried out to find an optimal EMM solution for highly inflected languages in general (See Steinberger et al. 2013).

Concretely, the Hungarian partners added 18 new news sources; they defined eighteen major news categories

Figure 3 Display in EMM-NewsBrief of part of a Hungarian news article together with the automatically extracted meta-information: News categories ('Political Unrest' and 'EC News') and related defining words found; entities recognised in the article and their mention frequency; quotation by and about a person (*Bailly* and *Yanukovich*, respectively). Note that the Hungarian spelling *Yanukovics* was automatically recognised as being equivalent to the EMM default spelling *Yanukovich*).

(using approximately 500 search words); they added 3215 Hungarian first names plus 2580 words to the NER trigger word lists (including titles, positions, professions, religions, nationalities, etc.); they verified over 60,000 geographical names to potentially add the Hungarian equivalent (e.g. *Bécs* for *Vienna*); they added over 100 reporting verb forms for quotation recognition (third person singular and plural, past and present tense of 36 lemmas), including those with separable prefix (e.g. *kijelent – jelent ki* – En: *declare*); they produced various stop word lists (generic ones, as well as lists to avoid recognition of wrong person or location names), and lists of modifiers and conjunctions, with altogether 1500 items. Finally, they provided nominal suffix lists containing the most frequently occurring simple suffix variants, 16 nominal case suffixes in their variant forms (eg. *nak/nek, on/en/ön/n, t/ot/et/öt/at*, etc.) and the plural ending (*k/ok/ek/ök/ak,* etc.) and possessive singular and plural forms (*e/ei/a/ai*), but not containing the theoretically possible complex inflected forms.

4.2 Quotation recognition

Quotations (reported speech) are recognised in EMM if the following three elements are found to be adjacent to each other in a text (see Pouliquen et al. 2007): (1) a person name (or a part of a name mentioned elsewhere in full), (2) a reporting verb (or a colon) and (3) text surrounded by any of a variety of quotation marks. Optionally, elements from a closed list of common modifiers (e.g. *on Tuesday*) can occur. References to other persons or organisations inside the quotation are also recognised, e.g.:

Merkel said on Tuesday "… Orbán …".

Figure 3 shows a screenshot from EMM-NewsBrief where the person quoting and a person mentioned in the quotation are displayed together with the quote and the reporting verb. EMM now recognises an average of 124 new quotes per day in an average of 2023 Hungarian news articles per day. An evaluation showed that, due to the strict adjacency requirement, precision is 100%, but we miss many quotes (in Hungarian as well as in other languages) because quotes are spread over several sentences and the name is not explicitly repeated. Some of these occurrences will be found once we have implemented Hungarian co-reference resolution, which includes common-noun co-reference such as *Victor Orbán* and *Fidesz leader*, exploiting the historically collected trigger words found next to each name (J. Steinberger et al. 2011). We are currently also testing a relaxation of the strict rules, allowing a small number of other words (excluding names) in between the three elements. A window of seven other words approximately doubles the number of recognised quotations, but it also leads to some recognition errors so that the currently running version of the quotation recognition in EMM still uses the strict adjacency requirement. The quotation recognition results are displayed as part of EMM's publicly accessible news analysis pages, i.e. together with the news clusters and also with each individual article in EMM-NewsBrief, but also as part of the entity pages in EMM-NewsExplorer.

4.3 Recognition and lemmatisation of names

Under the condition that no morphological analysis tools can be used in EMM, there are three main challenges to recognise named entities for languages with a productive morphology such as Hungarian (Steinberger et al. 2013): (a) the language-specific dictionaries (e.g. titles, nationalities and professions) are potentially very large and unforeseeable, due to the many inflection and agglutination forms; (b) the lookup of the known uninflected names will not work if these names are inflected in the text (EMM uses lists with over one million known names); (c) inflected newly recognised names need to be lemmatised in order to ground the name mention with a real-world entity and also to match other variants within the same and across different languages.

We experimented with three different solutions to these challenges: (1) ignore inflection and thus missing inflected word forms; (2) use wild cards for dictionaries and for the lookup procedure; (3) use hand-crafted suffix replacement rules that at least capture the most frequent inflection forms. These rules can be used to generate inflection forms of known names in order to improve the recall in the lookup procedure, or alternatively to lemmatise newly recognised, potentially inflected names.

Ignoring inflection (solution 1) leads to low recall: A small-scale manual evaluation showed that about 14% of

person names in the Hungarian news are inflected and a large-scale semi-automatic evaluation on one month of Hungarian news indicated that 39% of location name mentions are inflected. This number was derived by compiling frequency lists of names found in Hungarian text by looking for known place names with wild-cards and by then evaluating top-, middle- and low-frequency items.

In order to test lemmatisation (solution 3), we applied 66 hand-crafted suffix replacement rules to large lists of person and organisation names and we found that in about 80% of cases, name endings were cut off too generously because some inflection suffixes are also frequent name endings (e.g. the accusative marker *–t* for words ending in a vowel). In the next experiment, we only stripped off these suffixes when the resulting name was in our list of half a million previously identified known names. Using these conditions, 287 of 18414 names were lemmatised (1.56%), of which 7 wrongly (2.8%; e.g. the name *Gordon* was wrongly truncated to *Gord*). This method thus captured less than 12% of the expected name inflections (1.56% out of an expected 14% inflected names). For our purposes, the yield is too low considering the error rate.

When applying the wild card solution (solution 2) for the lookup of known names, it is important not to simply add the wild card at the end of the name because up to two final letters of the name may be replaced (e.g. *Obama* – *Obamával* – En: 'Obama–Obama.INS'). For four of the most frequent such suffix replacements, we thus considered this change (e.g. *–a* → *–á%*; i.e. searching for *Obamá%* to capture inflections of *Obama*), for person, organisation and location names. The results on a large set of Hungarian news articles showed that 32 out of 1025 identified distinct locations were wrong (3.12%), corresponding to 105 location *mentions* out of 4292 (2.45%). Out of these, 27 errors (69 occurrences, i.e. 67%) occurred because of wildcard usage. Many of the errors can be avoided by putting these words on a geo-stop word list, which we proceeded to do.

Wild cards were also applied to titles and other words used in the rules to recognise new names (classical NER) even if these occur in their inflected form (e.g. *neurobiokémikus%*, En: *neurobiochemist%*). An analysis of 1167 newly identified names (2136 mentions) in Hungarian news, using these wild card dictionaries, showed that 15.4% of these were inflected, confirming that the method did capture inflected forms very well. The recognition accuracy was 80%. An analysis of the major errors showed that they were not related to the usage of wild cards: (a) two names together were identified as one (30 names, 53 mentions); (b) only part of the name was recognised (14 names, 27 mentions); or (c) names got truncated due to a software bug (11 names, 29 mentions; this bug has since been fixed).

Interestingly, when looking at all names identified, the inflection rate is much higher for low-frequency names than for high-frequency names: 28% for frequency 1; 13.3% for frequency 5; 1.86% for the 1057 highest frequent names (corresponding to 913 inflected forms out of 48,938 name occurrences). It is not clear to us whether this

Table 1. Frequencies of suffixes, their standard forms and their grammatical function for names found in text with frequency 1, 5 and top-frequent (*Fq 1*, *Fq5* and *Fq TOP*, respectively), as well as in a set of names entirely newly recognised in Hungarian text (*Fq NR*) and in geographical location names (*Fq Geo*). The cells with the highest frequencies per category are shaded.

SUFFIX	Standard	Function	Fq 1	Fq 5	Fq TOP	Fq NR	Fq Geo
AKAT	K+T	PL+ACC				1	
A$	$	N->A Der				1	
AT	T	ACC	1	5		2	4
BA	BA	ILL	1	5		13	25
BAN	BAN	INE	1	35	187	45	498
BE	BA	ILL				1	13
BEN	BAN	INE		10	69	10	88
BÓL	BÓL	ELA		5		4	41
BŐL	BÓL	ELA				3	5
CSAL	VAL	INS			17		
CSEL	VAL	INS		5			
DAL	VAL	INS		5			
DEL	VAL	INS				2	
É	É	POS				6	4
ÉK	ÉK	Der			19	5	
ÉKKAL	ÉK+VAL	Der+INS		5			
ÉKBÓL	ÉK+BÓL	Der+ELA					
ÉKNAK	ÉK+NAK	Der+DAT					
ÉKON	ÉK+ON	Der+SUP					
ÉKRÓL	ÉK+RÓL	Der+DEL					
ÉKTŐL	ÉK+TŐL	Der+ABL					
EN	ON	SUP		20		8	300
ÉRT	ÉRT	CAU				5	9
ET	T	ACC	1	15	34	19	12
ETT		SUP					
FÉLE	FÉLE	N->A Der					
FAL	VAL	INS					
GAL	VAL	INS		5		6	9
GEL	VAL	INS		5			
GYEL	VAL	INS			20	1	
HEZ	HOZ	ALL				1	
HOZ	HOZ	ALL		5		9	13
IG	IG	TER					1
JÁT	JA+T	PERS+ACC				1	
JE	JA	PERS				1	
JÉT	JA+T	PERS+ACC				3	
KAL	VAL	INS				1	
KEL	VAL	INS					1
KÉNT	KÉNT	FOR		5			
LAL	VAL	INS				1	
LEL	VAL	INS		10		3	
LYAL	VAL	INS				2	
LYEL	VAL	INS					
MAL	VAL	INS		5			3
MEL	VAL	INS				1	
N	ON	SUP	1	5		13	79
NAK	NAK	DAT	3	40	76	54	31
NEK	NAK	DAT	1	40	179	39	4
NAL	VAL	INS		20		8	52
NÁL	NÁL	ADE		5		9	
NEL	VAL	INS	1	5		3	2
NÉL	NÁL	ADE	1	10		1	18
ON	ON	SUP	1	15		16	189
ŐN	ON	SUP					6
OT	T	ACC	1	15	24	20	41
ŐT	T	ACC		5	57		
RA	RA	SUB	1	5		34	6
RE	RA	SUB		5		2	26
REL	VAL	INS				5	
RAL	VAL	INS		5			
RÓL	RÓL	DEL		10		21	9
RŐL	RÓL	DEL		10		2	11
SAL	VAL	INS				1	
SZAL	VAL	INS				3	
SZEL	VAL	INS				1	
T	T	ACC	9	110	197	130	33
TAL	VAL	INS	1			3	
TEL	VAL	INS			19	2	
TÓL	TÓL	ABL	1	25		24	22
TŐL	TÓL	ABL	1			14	16
UK	JUK	PERS					
VAL	VAL	INS	2	30	15	13	7
VEL	VAL	INS		10		4	
ZEL	VAL	INS				1	
TOTAL			28	515	913	578	1578

corresponds to a morphological reality in Hungarian text (important people may be referred to mostly as actors, i.e. in the uninflected nominative case) or whether it is due to our way of recognising names. In order to get a better understanding of this data, we also evaluated separately names that had been newly recognised in Hungarian text and never seen before in other languages. This Hungarian sample consisted of 1167 name forms, equivalent to 2136 name mentions (type/token ratio is 1.83). The observed ratio of inflected name forms in this set was 15.4%. These numbers fit the assumption rather well that the lesser frequently names are more frequently inflected.

4.4 Hungarian name inflection pattern frequencies

As EMM has now been analysing Hungarian news for several months, we have gathered a lot of data on name inflections, especially on the frequency of inflection forms. High-, middle- and low-frequency lists of these have been analysed and categorised to identify the major suffix patterns for Hungarian and foreign names. **Table 1** shows the frequencies for 74 different inflection endings, observed with locations (Column *Fq Geo*) and with person names. For person names, we distinguish the frequencies with names found only once (*Fq 1* – the evaluated sample consisted of 100 names), with names found five times (*Fq 5* – 771 name forms were evaluated, equivalent to 3855 name occurrences) and with the names most frequently found (*Fq TOP* – 1057 name forms were evaluated, equivalent to 48,938 name occurrences). Furthermore, we separately looked at names newly recognised in Hungarian text, i.e. names that had not been found in other languages before (*Fq NR* – 1167 name forms were evaluated, equivalent to 2136 name mentions). These are thus names that are either relatively rarely mentioned, or names that may be frequently mentioned in Hungarian language news, but not so much in international news. The cells with the highest values in each of the columns are highlighted.

The analysis showed that over 70% of person name inflections are variant forms of accusative, dative and instrument cases, while over 70% of location inflections are inessive and superessive forms, corresponding to position 'inside a place' and position 'over a place', respectively. The majority of Hungarian city names take the superessive forms (e.g. *Budapest-en* – 'Budapest-SUP'), while all names of foreign cities take the inessive case (e.g. *Boston-ban* – 'Boston-INE') (see e.g. Megyesi 1998 for information on Hungarian grammar). We plan to make use of this detailed analysis in the future to improve how EMM handles Hungarian inflection. For instance, the frequency information could be used to lemmatise newly found person or location names, by stripping off only the most frequent suffixes (–ban, –nak, –nek, –t) and thus reducing the error rate for the lemmatisation process. The trickiest frequent inflection suffix to remove is the accusative ending –t as it consists of a single letter and there are many uninflected names ending in –t.

5. Conclusion

Dealing with a highly inflected language such as Hungarian, using up to 18 nominal cases, vowel harmony and agglutination without using a morphological analyser seems to be an almost insurmountable challenge. However, experiments testing various shallow methods and tuning them yielded very positive and encouraging results. We have not currently found a way to lemmatise newly found inflected names, but we have identified a viable practical solution: As the uninflected nominative form is by far the most frequent, this form will be the first to become a 'known name' in EMM (i.e. having been found at least once each in at least five different news clusters). Once a name has acquired the status of *known name*, the system will automatically also search for its inflection forms, using the wild card so that, from that moment on, inflection forms of that known name will also be recognised.

It is our plan to implement this same pragmatic solution for other highly inflected languages, including those of the Slavic and other languages of the Finno-Ugric family (Steinberger et al. 2013). Using this method to treat inflection in each highly inflected language will thus mostly require (a) confirming that the uninflected nominative case is indeed the most frequent name form; (b) identifying where to add wild cards for the lookup of trigger words (e.g. titles) and of known names (persons, organisations and locations); (c) manually evaluating large frequency lists of names recognised with these wild cards; (d) possibly tune the wild card patterns; (e) produce stop word lists of other words that were erroneously picked up by the wild card patterns and (f) evaluate the NER results. The required effort is manageable and entirely within the limits of what the EMM team finds acceptable when adding a new language to its media monitoring tool set.

The publicly accessible EMM webpages accessible via http://emm.newsbrief.eu/overview.html show the live results of this collaborative effort between the JRC and the Hungarian Academy of Sciences.

6. References

Cohen W. William and Sunita Sarawagi (2004). Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods. In Conference on Knowledge Discovery in Data: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 89–98.

Farkas Richárd, Veronika Vincze, István Nagy, Róbert Ormándi, György Szarvas, and Attila Almási (2008). Web-Based Lemmatisation of Named Entities. In Proceedings of the 11th international conference on Text, Speech and Dialogue, TSD '08, Springer-Verlag: Berlin, Heidelberg,. pp. 53–60.

Gábor Kata, Enikő Héja, Ágnes Mészáros, and Bálint Sass, (2003). Nyílt tokenosztályok reprezentációjának technológiája [En: Representation of open token classes]. Szeged. IKTA-00037/2002, interim report.

McDonald D. David (1996). Internal and External Evidence in the Identification and Semantic

Categorization of Proper Names. In Branimir Boguraev and James Pustejovsky (eds.): Corpus Processing for Lexical Acquisition, MIT Press: Cambridge, MA, pp. 21–39.

Megyesi Beáta (1998). The Hungarian Language. A Short Descriptive Grammar http://stp.lingfil.uu.se/~bea/publ/megyesi-hungarian.pdf, University of Uppsala.

Novák Attila (2003). Milyen a jó Humor? [En: What is good Humour like?] In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, pp. 138-145.

Piskorski Jakub, Karol Wieloch and Marcin Sydow (2009). On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages. Inf. Retrieval (2009) 12:275–299.

Piskorski Jakub & Marcin Sydow (2007). Usability of String Distance Metrics for Name Matching Tasks in Polish. In: Proceedings of the 3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, (LTC'2007), Poznań, Poland, 5-7.10.2007.

Pouliquen Bruno and Ralf Steinberger (2009). Automatic Construction of Multilingual Name Dictionaries. In: Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster (eds.): Learning Machine Translation. pp. 59-78. MIT Press - Advances in Neural Information Processing Systems Series (NIPS).

Pouliquen Bruno, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, Clive Best (2006). Geocoding multilingual texts: Recognition, Disambiguation and Visualisation. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 53-58. Genoa, Italy, 24-26 May 2006.

Pouliquen Bruno, Ralf Steinberger, and Clive Best (2007). Automatic Detection of Quotations in Multilingual News**.** In: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'2007), pp. 487-492. Borovets, Bulgaria, 27-29.09.2007.

Simon Eszter (2013). Approaches to Hungarian Named Entity Recognition. PhD Thesis. Budapest University of Technology and Economics, Budapest.

Steinberger Josef, Jenya Belyaeva, Jonathan Crawley, Leonida Della Rocca, Mohamed Ebrahim, Maud Ehrmann, Mijail Kabadjov, Ralf Steinberger, and Erik van der Goot (2011). Highly Multilingual Coreference Resolution Exploiting a Mature Entity Repository. Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP'2011), pp. 254-260. Hissar, Bulgaria, 12-14 September 2011.

Steinberger Ralf (2013). Multilingual and cross-lingual news analysis in the Europe Media Monitor (EMM)**.** In: Mihai Lupu, Evangelos Kanoulas, and Fernando Loizides (eds.): *Multidisciplinary Information Retrieval.* 6th Information Retrieval Facility Conference (IRFC'2013), Limassol, Cyprus. Springer Lecture Notes in Computer Science, Vol. 8201, pp. 1-4.

Steinberger Ralf, Bruno Pouliquen, and Erik van der Goot (2009). An Introduction to the Europe Media Monitor Family of Applications. In: Fredric Gey, Noriko Kando, and Jussi Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), pp. 1-8. Boston, USA. 23 July 2009.

Steinberger Ralf, Maud Ehrmann, Júlia Pajzs, Mohamed Ebrahim, Josef Steinberger, and Marco Turchi (2013). Multilingual media monitoring and text analysis - Challenges for highly inflected languages. In: Ivan Habernal and Václav Matoušek (eds.). Text, Speech and Dialogue. 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 2013, Proceedings. Springer Lecture Notes in Artificial Intelligence LNAI 8082, pp. 22-33.

Steinberger Ralf, Sylvia Ombuya, Mijail Kabadjov, Bruno Pouliquen, Leonida Della Rocca, Jenya Belyaeva, Monica De Paola, and Erik van der Goot (2011). Expanding a multilingual media monitoring and information extraction tool to a new language: Swahili. Language Resources and Evaluation Journal (DOI 10.1007/s10579-011-9165-9), Volume 45, Issue 3, pp. 311-330.

Szarvas György, Richárd Farkas, and András Kocsor (2006b). A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In Proceedings of Discovery Science 2006, pp. 267–278. Springer Verlag.

Szarvas György, Richárd Farkas, László Felföldi, András Kocsor, and János Csirik (2006a). A highly accurate Named Entity corpus for Hungarian. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06), pp. 1957-1960.

Trón Viktor, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga (2005). HunMorph: open source word analysis. In: Proceedings of the ACL 2005 Software Workshop, pp. 77-85.

Tsuruoka Yoshimasa and Jun'ichi Tsujii (2003). Boosting Precision and Recall of Dictionary-Based Protein Name Recognition. In Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, pp. 41–48, Sapporo, Japan. Association for Computational Linguistics.

Varga Dániel and Simon Eszter (2007). Hungarian named entity recognition with a maximum entropy approach. Acta Cybernetica, 18:293–301.

Zaghouani Wajdi, Bruno Pouliquen, Mohamed Ebrahim, and Ralf Steinberger (2010). Adapting a resource-light highly multilingual Named Entity Recognition system to Arabic. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 563-567. Valletta, Malta, 19-21 May 2010.