K + K = 120: Papers dedicated to L. Kálmán & A. Kornai on the occasion of their 60th birthdays, pp. 481–497

The definition of Named Entities

ESZTER SIMON
Research Institute for Linguistics,
Hungarian Academy of Sciences
simon.eszter@nytud.mta.hu

KEYWORDS

computational linguistics
named entity recognition
annotation schemes
proper names
compositionality

ABSTRACT

Named Entity Recognition (NER) is one of the most intensively studied tasks of computational linguistics. It has two substeps: first, locating the Named Entities (NEs) in unstructured texts, and second, classifying them into pre-defined categories. A key issue is how to define NEs. This issue interconnects with the issue of selection of classes and the annotation schemes applied in the field of NER. The major standard guidelines do not give an exact definition of NEs, but rather list examples and counterexamples. For getting a usable definition of NEs, we investigate the approach taken in the philosophy of language and linguistics, and we map our findings to the NER task. We do not wish to give a complete description of the theory and typology of proper names but to find a plausible way to define linguistic units relevant to the NER task.

1. Introduction

Named Entity Recognition (NER), the task of automatic identification of selected types of Named Entities (NEs), is one of the most intensively studied tasks of Information Extraction (IE). Presentations of language analysis typically begin by looking words up in a dictionary and identifying them as nouns, verbs, adjectives, etc. But most texts include lots of names, and if a system cannot find them in the dictionary, it cannot identify them, making it hard to produce a linguistic analysis of the text. Thus, NER is of key importance in many Natural Language Processing (NLP) tasks, such as Information Retrieval (IR) or Machine Translation (MT).

The NER task, which is often called Named Entity Recognition and Classification in the literature, has two substeps: first, locating the NEs in unstructured texts, and second, classifying them into pre-defined categories. A key issue is how to define NEs. This issue interconnects with the

issue of selection of classes and the annotation schemes applied in the field of NER.

The major standard guidelines applied in the field of NER do not give an exact definition of NEs, but rather list examples and counterexamples. The only common statement they make is that NEs have unique references. For getting a usable definition of NEs, we investigate the approach taken in the philosophy of language and linguistics, and we map our findings onto the NER task. We do not wish to give a complete description of the theory and typology of proper names, but to find a plausible way to define linguistic units relevant for the NER task.

The article is structured as follows.[1] In section 2, we give an overview of the annotation schemes applied in the field of NER. Section 3 describes the philosophical approach, and section 4 gives the linguistic background of the theory of proper names. The article concludes in section 5 with the most important findings about mapping the theory of proper names to the NER task.

2. Annotation schemes

2.1. MUCs

The first major event dedicated to the NER task was the 6th Message Understanding Conference *(MUC-6)* in 1995. As the organizers write in their survey about the history of MUCs (Grishman & Sundheim 1996), these conferences were rather similar to shared tasks, because the submission of participants' results was a prerequisite for participation at the conference. Prior MUCs focused on other IE tasks; MUC-6 was the first including the NER task, which consisted of three subtasks (Sundheim 1995):

– entity names (ENAMEX): organizations, persons, locations;
– temporal expressions (TIMEX): dates, times;
– number expressions (NUMEX): monetary values, percentages.

The annotation guidelines define NEs as "unique identifiers" of entities, and give an enormous list of what to annotate as NEs. However, the best support for annotators is the restriction about what not to annotate: "names that do not identify a single, unique entity".

[1] This article is a slightly modified version of a chapter of the author's PhD dissertation (Simon 2013).

As for the temporal expressions, the guidelines distinguish between absolute and relative time expressions. To be considered absolute, the expression must indicate a specific segment of time, e.g.,

(1) twelve o'clock noon

(2) January 1979

A relative time expression indicates a date relative to the date of the document, or a portion of a temporal unit relative to the given temporal unit, e.g.,

(3) last night

(4) yesterday evening

In MUC-6, only absolute time expressions were to be annotated.

The numeric expressions subsume monetary and percentage values. Modifiers that indicate the approximate value of a number are to be excluded from annotation, e.g.,

(5) about 5%

(6) over $90,000

The modified version of the MUC-6 guidelines was used for the *MUC-7* NER task in 1998 (Chinchor 1998). The most notable change was that relative time expressions became taggable. The MUC-7 guidelines became one of the most widely used standards in the field of NER. They were used with slight modifications for the Multilingual Entity Tasks (MET-1 and 2) (Merchant et al. 1996) and for the Hub-4 Broadcast News Evaluation (Miller et al. 1999) in 1999.

According to the MUC guidelines, embedded NEs can also be annotated, e.g.,

(7) The [morning after the [July 17]$_{\text{DATE}}$ disaster]$_{\text{TIME}}$

2.2. CoNLL

The Computational Natural Language Learning *(CoNLL)* conference is the yearly meeting of the Special Interest Group on Natural Language Learning (SIGNLL) of the Association for Computational Linguistics (ACL). Shared tasks organized in 2002 and 2003 were concerned with language-independent NER (Tjong Kim Sang 2002; Tjong Kim Sang & De Meulder 2003). Annotation guidelines were based on the NER task definition of the MITRE Corporation (http://www.mitre.org/) and the Science Applications International Corporation (SAIC) (Chinchor et al. 1999), which are slightly modified versions of the MUC guidelines. A new type, `Measure`, was introduced for NUMEX elements, e.g.,

(8) 23 degrees Celsius

In contrast to the MUC guidelines, instructions are given regarding certain kinds of metonymic proper names, decomposable and non-decomposable names, and miscellaneous non-taggables. The latter constitute a new category, `Miscellaneous`, which includes names falling outside the classic ENAMEX, e.g., compounds that are made up of locations, organizations, etc., adjectives and other words derived from a NE, religions, political ideologies, nationalities, or languages.

2.3. ACE

As part of the *Automatic Content Extraction* (ACE) program (a series of IE technology evaluations from 1999 organized by the National Institute of Standards and Technology (NIST)), new NE types were introduced in addition to the classic ENAMEX categories: `Facility`, `Geo-Political Entity`, `Vehicle` and `Weapon`. The category `Facility` subsumes artifacts falling under the domains of architecture and civil engineering. `Geo-Political Entities` are composite entities comprised of a population, a government, a physical location, and a nation (or province, state, county, city, etc.). The seven main types are divided into dozens of subtypes and hundreds of classes (ACE 2008) . The ACE program is concerned with automatic extraction of content, including not only NEs but also their relationships to each other and events concerning them. For the purposes of this more complex task, all references to entities are annotated: names, common nouns, noun phrases, and pronouns. In this regard, ACE is exceptional in the race of NER standards, where common nouns and pronouns are not to be annotated.

2.4. LDC

The Linguistic Data Consortium (LDC) has developed annotation guidelines for NEs and time expressions within the *Less Commonly Taught Languages* (LCTL) project. In contrast to the ones mentioned above, these guidelines give an exact definition of NEs (LDC 2006) : "An entity is some object in the world – for instance, a place or a person. A named entity is a phrase that uniquely refers to that object by its proper name, acronym, nickname or abbreviation." Besides the classical name categories (`PER, ORG, LOC`), they also annotate `Titles`, which are separated from the person's name, e.g.,

(9) said $[\text{GlobalCorp}]_{\text{ORG}}$ $[\text{Vice President}]_{\text{TTL}}$ $[\text{John Smith}]_{\text{PER}}$

The LCTL annotation guidelines are the first concerned with meaning and compositionality of NEs: "The meaning of the parts of names are not typically part of the meaning of the name (i.e., names are not *compositional*) and, therefore, names cannot be broken down into smaller parts for annotation." Thus, a NE is treated as an indivisible syntactic unit that cannot be interrupted by an outside element.

In addition to the classical ENAMEX, TIMEX and NUMEX categories, there are a wide range of other, marginal types of NEs, which are relevant for particular tasks, e.g., extracting chemical and drug names from chemistry articles (Krallinger et al. 2015); names of proteins, species, and genes from biology articles (Ding et al. 2015); or project names, email addresses and phone numbers from websites (Zhu et al. 2005).

2.5. Summary

Early works define the NER problem as the recognition of proper names in general. Names of persons, locations and organizations have been studied the most. Besides these classical categories, there is a general agreement in the NER community about the inclusion of temporal expressions and some numerical expressions, such as amounts of money and other types of units. The main categories can be divided into fine-grained subtypes and classes, and marginal types are sometimes included for specific tasks. Annotation guidelines usually do not go further in defining NEs than saying that they are "unique identifiers" or that they "uniquely refer" to an entity. Only one of the guidelines mentions the meaning and compositionality of NEs: it postulates NEs as indivisible units, although earlier guidelines allow embedded NEs.

3. Language philosophical views: from Mill to Kripke

3.1. John Stuart Mill

"A proper name is a word that answers the purpose of showing what thing it is that we are talking about, but not of telling anything about it", writes John Stuart Mill in his 1843 *A system of logic* (Mill 2002). According to him, the semantic contribution of a name is its referent and only its referent. One of his examples illustrating this statement is the name of the town Dartmouth. The town was probably named after its localization, because it lies at the mouth of the river Dart. But if the river had changed its course, so that the town no longer lay at the mouth of the Dart, one could still use the name *Dartmouth* to refer to the same place as before. Thus, it is not part of the meaning of the name *Dartmouth* that the town with this name lies at the mouth of the Dart.

3.2. Gottlob Frege and Bertrand Russell

Gottlob Frege's puzzle of the Morning Star and the Evening Star challenges the Millian conception of names. In his famous work *Über Sinn und Bedeutung* (Frege 2000), he distinguishes between *sense (Sinn)* and *reference (Bedeutung)*. Without the distinction between sense and reference, the following sentences would be equal:

(10) The Morning Star is the Evening Star.

(11) The Morning Star is the Morning Star.

Both names have the same reference (Venus), so they should be interchangeable. However, since the thought expressed by (10) is distinct from the thought expressed by (11), the senses of the two names are different. While (11) seems to be an empty tautology, (10) can be an informative statement, even a scientific discovery. If somebody did not know that the Evening Star is the Morning Star, he/she could think that (11) was true, while (10) was false.

To solve the puzzle, without resorting to a two-tiered semantic theory, Bertrand Russell used the description theory. The *description theory of names* states that each name has the semantic value of some definite description (Cumming 2012). For example, *Aristotle* might have the semantic value of 'the teacher of Alexander the Great'. *The Morning Star* and *the Evening Star* might correspond to different definite descriptions

in their semantic value, and would make different semantic contributions to the sentences in which the occur.

Frege and Russell both argue that Mill was wrong: a proper name is a definite description abbreviated or disguised, and such a description gives the sense of the name. According to Frege, a description may be used synonymously with a name, or it may be used to fix its reference.

3.3. Saul Kripke

Saul Kripke concurred only partially with Frege's theory. Description fixes reference, but the name denoting that object is then used to refer to that object, even if referring to counterfactual situations where the object does not have the properties in question, writes Kripke in *Naming and necessity* (Kripke 1981). One of Kripke's examples is Gödel and the proof of incompleteness of arithmetic. If it turned out that Gödel was not the man who proved the incompleteness of arithmetic, Gödel would not be called 'the man who proved the incompleteness of arithmetic', but he would still be called 'Gödel'. Thus, names are not equal to definite descriptions.

Kripke postulates proper names as *rigid designator*s. Something is a rigid designator if it designates the same object in every possible world. The concept of a possible world (or counterfactual situation) is used in modal semantics, where the sentence *Frank might have been a revolutionist* is interpreted as a quantification over possible worlds. Kripke suggests an intuitive test to find out what is a rigid designator. An updated example: *the President of the US in 2017* designates a certain man, Trump; but someone else (e.g., Clinton) may have been the President in 2017, and Trump might not have; so this designator is not rigid. When talking about what would happen to Trump in a certain counterfactual situation, we are talking about what would happen to *him*. So 'Trump' is a rigid designator.

With respect to proper names, reference can be fixed in various ways. In the case of initial baptism it is typically fixed by ostension or description. Otherwise, the reference is usually determined by a chain, passing the name from link to link. In general, the reference depends not just on what we think, but on other people in the community, the history of how knowledge of the name has spread. It is by following a history that one gets to the reference.

Kripke argues that proper names are not the only kinds of rigid designators: species names, such as *tiger*, or mass terms, such as *gold*, certain terms for natural phenomena, such as *heat*, and measurement units, such as *one meter* are further examples. There is a difference between the phrase

one meter and the phrase *the length of the metre bar at* t_0. The first phrase is meant to designate rigidly a certain length in all possible worlds, which in the actual world happens to be the length of the metre bar at t_0. On the other hand, *the length of the metre bar at* t_0 does not designate anything rigidly.

3.4. Summary

Kripke goes back to the Millian theory of names, and at the same time breaks with Frege's theory, when he writes that proper names do not have sense, only reference. He declares that a proper name is a rigid designator, which designates the same object in every possible world. Through examples he proves that definite descriptions are not synonymous with names, but they can still fix a referent. In the case of proper names, the reference can be fixed in an initial baptism, after which the name spreads in the community by a chain, from link to link. In Kripke's theory, species names, mass terms, natural phenomena and measurement units are also rigid designators.

4. The linguistic approach

Besides the theory of rigid designators, another concept used in the literature to define NEs is that of unique reference. In subsection 4.1, we clarify the meaning of the phrase "unique reference", which seems to be used non-systematically in NER guidelines. Unique reference can act as the separator line between proper names and common nouns. There are however certain *linguistic properties* by which we can make a stronger distinction, as described in subsection 4.2. The main feature distinguishing between them is the issue of compositionality, which is discussed in subsection 4.3. Finally, we sum up our findings about the linguistic background of proper names in subsection 4.4.

4.1. Unique reference

In the MUC guidelines (Chinchor 1998), the definition of what to annotate as NEs is as follows: "proper names, acronyms, and perhaps miscellaneous other unique identifiers", and what not to annotate as NEs: "artifacts, other products, and plural names that do not identify a single, unique entity".

In the LCTL guidelines we find the following definition: "a NE is a phrase that uniquely refers to an object by its proper name, acronym, nickname or abbreviation" (LDC 2006).

Let us take these definitions one by one. In the first case, the phrase "unique identifiers" is coordinated with "proper names" and "acronyms", and "unique" is an attributive adjective modifying the noun "identifiers". Thus, "unique" means here that the identifier is unique, similarly to proper names and acronyms. In the second case, however, it is the entity a linguistic unit refers to that must be unique in order for the unit to qualify as a NE. In the LCTL guidelines, the phrase "uniquely refers" means something similar as in the first case, it is therefore the referring linguistic unit that must be unique, not the entity in the world to which it refers.

Here and in several other places in the literature, the difference between the concepts of referring act and reference seems to be blurred. When trying to determine what is unique, we find that in most grammar books the names and the entities they refer to are not clearly distinguished. However, it does matter whether we are talking about Charlie or about the name *Charlie*. To prevent such an ambiguity, we always indicate the meta-linguistic usage by single quotation marks.

By investigating various definitions of proper names, we can conclude that names refer to a unique entity (e.g., *London*), so names have unique reference (Quirk & Greenbaum 1980), in contrast to common nouns, which refer to a class of entities (e.g., *cities*), or non-unique instances of a certain class (e.g., *city*). However, we can refer to and even identify an entity by means of common nouns. The difference is that proper names, even standing by themselves, always identify entities, while a common noun can do so only in such cases when it constitutes a noun phrase with other linguistic units. Common nouns may stand with a possessive determiner (e.g., *my car*), or with a demonstrative (e.g., *this car*), or can be a part of a description (e.g., *the car that I saw yesterday*).

Many proper names share the feature of having only one possible reference, but a wide range of them refer to more than one object in the world. For example, *Washington* can refer to thousands of people who have *Washington* as their surname or given name, a US state, the capital of the US, cities and other places throughout America and the UK, roads, lakes, mountains, educational organizations, and so forth. These kinds of proper names are referentially multivalent (Anderson 2007), but each of the references is still unique.

Some proper names occur in plural form, optionally or exclusively. In the latter case, the plural suffix is an inherent part of the name. These are

the so called *pluralia tantum* (e.g., *Carpathians, Pleiades*). According to their surface form, it might seem that they can be broken down into smaller pieces, but the Carpathians do not consist of $carpathian_1$, $carpathian_2$, ..., $carpathian_n$, just as the Pleiades do not consist of pleiades. These names refer to groups of entities considered unique.

Names of brands, artifacts, and other products can be optionally used in plural form. For example, *Volvo* is a proper name referring to a unique company. But if we put it in a sentence, like *He likes Volvos*, it will refer to particular vehicles. This is a kind of metonymy, with the company name used to refer to a product of this company. Proper names in plural form can also be used in other kinds of figures of speech, for example in metaphors. In the phrase *a few would-be Napoleons*, some characteristics of the emperor are associated with men to which the word *Napoleons* refers. In these cases, proper names act like common nouns, i.e., they have no unique reference.

Additionally, there is a quite large number of linguistic units which are on the border between proper names and common nouns, because it is difficult to determine whether their reference is unique. Typically, they are used as proper names in some languages, but as common nouns in other ones. The difficulty of classification is usually mirrorred even in the spelling rules. For example, in the case of events (*World War II, Olympic Games* in English; *2. világháború, olimpiai játékok* in Hungarian; *Segunda Guerra Mundial, Juegos Olímpicos* in Spanish; *Seconde Guerre mondiale, Jeux olympiques* in French), expressions for days of the week and months of the year (*Monday, August* in English; *hétfő, augusztus* in Hungarian; *lunes, agosto* in Spanish; *lundi, août* in French), expressions for languages, nationalities, religions and political ideologies (*Hungarian, Catholic, Marxist* in English; *magyar, katolikus, marxista* in Hungarian; *húngaro, católica, marxista* in Spanish; *hongrois, catholique, marxiste* in French), etc. Categories vary across languages, so there seems to be no language-independent, general rule for classifying proper names.

4.2. Distinction between proper names and common noun phrases

As mentioned above, proper nouns are distinguished from common nouns on the basis of the uniqueness of their reference. However, we can make a stronger distinction based on other linguistic properties.

First, we have to clarify the disctinction between proper nouns and proper names made by current works in linguistics (e.g., Anderson 2007; Huddleston & Pullum 2002). Since the term "noun" is used for a class of single words, only single-word proper names are proper nouns: *Ivan* is both

a proper noun and a proper name, but *Ivan the Terrible* is a proper name that is not a proper noun. From this distinction follows that proper names cannot be compared to a single common noun, but to a noun phrase headed by a common noun. A proper noun by itself constitutes a noun phrase, while common nouns need other elements. In subsection 4.1, we gave a few examples. In the subsequent analysis, proper names and common noun phrases are juxtaposed.

Distinction between proper nouns and common nouns is commonly made with reference to *semantic properties*. One of them is the classic approach: entities described by a common noun, e.g., *horse*, are bound together by some resemblances, which can be summed up in the abstract notion of 'horsiness' or 'horsehood' (Gardiner 1957). A proper name, on the contrary, is a distinctive badge: there is no corresponding resemblance among the Charlies that could be summed up as 'Charlieness' or 'Charliehood'. Thus, we can say that common nouns realize abstraction, while proper names make distinction. However, Katz (1972) argues that the meaninglessness of names means that one cannot establish a semantic distinction between proper names and common noun phrases. The latter are compositional, because their meaning is determined by their structure and the meanings of their constituents (Szabó 2008), while proper names "allow no analysis and consequently no interpretation of their elements", quoting Saussure (1959). Thus, proper names are arbitrary linguistic units, and are therefore not compositional (see 4.3 for more details).

Moving on to *syntax*, common noun phrases are compositional, i.e., they can be divided into smaller units, while proper names are indivisible syntactic units. This is confirmed by the fact that proper names – as opposed to common nouns – cannot be modified internally, as can be seen in these examples:

(12) my son's college

(13) my son's beautiful college

(14) beautiful King's College

(15) *King's beautiful College

Further evidence is that in Hungarian and other highly agglutinative languages, the inflection always goes to the end of the proper name constituting a noun phrase. (16) presents the inflection of a proper name (here: a title), while (17) shows its common noun phrase counterpart (consider the second determiner in the latter):

(16) Láttam az Egerek és embereket.
'I saw (Of Mice and Men).ACC'

(17) Láttam az egereket és az embereket.
'I saw the mice.ACC and the men.ACC'

From the perspective of *morphology*, proper names must always be sacred, which means that the original form of a proper name must be reconstructible from the inflected form (Deme 1956). This requirement is mirrorred even in the current spelling rules in Hungarian: e.g., *Papp-pal* 'with Papp', *Hermann-nak* 'to Hermann'. Some proper names in Hungarian have common noun counterparts, as well, e.g., *Fodor* ∼ *fodor* 'frill', *Arany* ∼ *arany* 'gold'. Since the word *fodor* is exceptional, when inflecting it as a common noun, the rule of vowel drop is applied: *fodrot* 'frill.ACC'. However, when inflecting it as a proper name, it is inflected regularly, without dropping the vowel: *Fodort* 'Fodor.ACC'. The common noun *arany* also has exceptional marking, it is lowering, which means that it has *a* as a link vowel in certain inflectional forms, e.g., in the accusative, instead of the regular bare accusative marker: *arany-at* 'gold-ACC'. But as a proper name, it is inflected regularly: *Arany-t* 'Arany-ACC' (for more details, see Kornai 1994 and Kenesei et al. 1998). Psycholinguistic experiments on Hungarian morphology also confirm that proper names are inflected regularly (Lukács 2001), while common nouns may have exceptional markings.

4.3. The non-compositionality of proper names

In order to examine whether proper names are compositional or arbitrary linguistic units, here we give an analysis of how knowledge about the named entity can be deduced from the name. Proper names are not simply arbitrary linguistic units, but they show the arbitrariness most clearly of all, since one can give any name to his/her dog, ship, etc. It follows from the arbitrariness of the initial baptism that proper names say nothing about the properties of the named entity, in fact they do not even indicate what kind of entity we are talking about (a dog, a ship, etc.).

Although *monomorphemic* proper names are classic examples of non-compositionality, they are not semantically empty. For instance, Charlie is a boy by default, but this name is often given to girls in the US, and of course it can be given to pets or products. Semantic implications of proper names (if any) are therefore defeasible. This is in contrast with common nouns, since we cannot call a table 'chair' without violating the

Gricean maxims (Grice 1975). Monomorphemic proper names have only one non-defeasible semantic implication, namely if one is called *X*, then the predicate 'it is called *X*' will be true (cf. the Millian theory of proper names in section 3).

In the context of the current analysis, two types of *polymorphemic* proper names can be distinguished. First, there are phrases which are headed by a common noun and modified by a proper name, e.g., *Roosevelt square, Columbo pub*. The second type consists of two (or more) proper nouns, e.g., *Theodore Roosevelt, Volvo S70*.

In the case of the former, more frequent type, every non-defeasible semantic implication (except the fact of the naming) comes from the head, the modifier does not make any contribution. This can be shown by removing the head: from the sentence *You are called from the Roosevelt*, one cannot determine the source of the call, which might come from the Roosevelt Hotel, from the Roosevelt College, or from a bar in Roosevelt square. All we have is the trivial implication, that Roosevelt is the name of the place. The fact that the modifier contributes nothing to the semantics of the entire construction can be illustrated better by replacing the proper names with empty elements, e.g., *A square, B pub*. The acceptability of the construction is not compromised even in this case. One further argument against compositionality is that if we try to apply it to polymorphemic proper names, we get unacceptable result: Roosevelt has not lived at Roosevelt square, and Columbo has never been to the Columbo pub.

In the second construction, both head and modifier are proper nouns. The only contribution made by the head to the semantics of the phrase is that we know that the thing referred to by the modifier is a member of the group of things referred to by the head, e.g., Volvo S70 is a kind of Volvo, but not a kind of S70.

Regarding polymorphemic proper names in general, we can say that the head *H* bears the semantics of the entire construction, while the only contribution of the modifier *M* is that it shows that *M* is called 'M' and that it is a kind of *H*. This is in contrast with the classic compositional semantics of common nouns, where the *red hat* means a hat which is red, the former president used to be a president, etc., and these implications are non-defeasible.

4.4. Summary

This section gives an overview how we can distinguish between proper names and common nouns using an approach based in linguistics. The first distinguishing property is the unique reference: common nouns, standing by themselves, never have unique reference. They have to be surrounded by other constituents within a phrase to refer some unique entity in the world, while proper nouns have unique reference on their own. There are, however, proper names which seemingly refer to several entities; it is shown through examples that these do have unique reference. Additional linguistic properties of proper names are presented, based on which a stronger distinction between proper names and common nouns can be made. The distinction based on semantic properties is the clearest: common noun phrases are compositional while proper names are not.

5. Conclusion

As can be seen from this overview, the definition of proper names is still an open question in both philosophy and linguistics. If we try to apply the findings presented above to the NER task, we will face various challenges. However, there are a few statements which can be used as pillars of defining what to annotate as NEs.

Early works formulated the NER task as recognizing proper names in general. This generality posed a wide range of problems, so the domain of units to be annotated as NEs had to be restricted. In this restricted domain, we only find person and place names, which have been postulated as proper names from the very beginnings of linguistics (e.g., in Plato's dialogue, *Cratylus*, and in Dionysius Thrax' grammar). The third classical name type, the type of organization names has been mentioned in grammar books from the 19th century. Although the range of linguistic units to annotate was cut, the challenges have remained, since these kinds of names already exhibit properties which make the NER task difficult.

In the expression "named entity", the word "named" aims to restrict the task to only those entities where rigid designators stand for the reference (Nadeau & Sekine 2007). Something is a rigid designator if it designates the same object in every possible world and thus has unique reference – unique in every possible world. Rigid designators include proper names as well as species names, mass terms, natural phenomena and measurement units. These natural kind terms are only partially included in the NER task. The MUC guidelines allow for annotating measures (e.g., *16 tons*) and

monetary values (e.g., *100 dollars*), which are rigid designators according to Kripke's theory. Some temporal expressions, typically absolute time expressions, are also rigid designators (e.g., *the year 2017* is the 2017th year of the Gregorian calendar), but there are also many non-rigid ones, typically the relative time expressions (e.g., *June* is a month of an undefined year). Thus, the rigid designator theory must be restricted to keep out species names, mass terms and certain natural phenomena, but must also be loosened to allow tagging relative time expressions as NEs.

If we say that every linguistic unit which has unique reference must be annotated as a NE, we should annotate common noun phrases as well. However, dealing with common nouns is not part of the NER task, so other linguistic properties of proper names and common nouns must be considered to make the distinction between them stronger. The greatest difference is the issue of compositionality. Applying Mill's, Saussure's, and Kripke's theory about the meaninglessness of names, we must conclude that proper names are arbitrary linguistic units, whose only semantic implication is the fact of the naming. Thus, the semantics of proper names is in total contrast with the classic compositional semantics of common nouns, as they are indivisible and non-compositional units. To map it to the NER task: embedded NEs are not allowed, and the longest sequences must be annotated as NEs (e.g., in the place name *Roosevelt square* there is no person name 'Roosevelt' annotated).

There still remain a quite large number of linguistic units which are difficult to categorize. Typically, they are on the border between proper names and common nouns, which is confirmed by the fact that their status varies across languages. We should not forget that the central aim of the NER task is extracting important information from raw text, most of which is contained by NEs. Guidelines should be flexible enough to allow the annotation of such important pieces of information. For getting a usable definition of NEs, the classic Aristotelian view on classification, which states that there must be a *differentia specifica* which allows something to be the member of a group, and excludes others, is not applicable. For our purposes, the prototype theory (Rosch 1973) seems more plausible, where proper names form a continuum ranging from prototypical (person and place names) to non-prototypical categories (product and language names; Langendonck 2007 – consider the parallelism with the order in which names are mentioned in grammar books). Finally, the goal of the NER application will further restrict the range of linguistic units to be taken into account.

References

ACE. 2008. ACE (Automatic Content Extraction) English annotation guidelines for Entities. Version 6.6. Linguistic Data Consortium.

Anderson, J. M. 2007. The grammar of names. Oxford: Oxford University Press.

Chinchor, N. 1998. MUC-7 Named Entity Task Definition version 3.5. In Proceedings of the 7th Message Understanding Conference (MUC-7).

Chinchor, N., E. Brown, L. Ferro and P. Robinson. 1999. Named Entity Recognition Task definition version 1.4.

Cumming, S. 2012. Names. In E. N. Zalta (ed.) The Stanford encyclopedia of philosophy (Winter 2012 edition). Stanford, CA: The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University.

Deme, L. 1956. Családneveink alaki sérthetetlenségéről [On the morphological sacredness of Hungarian family names]. Magyar Nyelv 52. 365–368.

Ding, R., C. N. Arighi, J.-Y. Lee, C. H. Wu and K. Vijay-Shanker. 2015. pGenN, a gene normalization tool for plant genes and proteins in scientific literature. PLoS ONE 10.

Frege, G. 2000. Ueber Sinn und Bedeutung. In Stainton (2000, 45–64).

Gardiner, A. 1957. The theory of proper names. A controversial essay. Oxford: Oxford University Press.

Grice, H. P. 1975. Logic and conversation. In P. Cole and J. L. Morgan (eds.) Syntax and semantics, vol. 3: Speech acts. New York: Academic Press. 41–58.

Grishman, R. and B. Sundheim. 1996. Message Understanding Conference – 6: A brief history. In Proceedings of the 16th International Conference on Computational Linguistics (COLING). Kopenhagen, 466–471.

Huddleston, R. and G. K. Pullum. 2002. The Cambridge grammar of the English language. Cambridge: Cambridge University Press.

Katz, J. J. 1972. Semantic theory. New York: Harper and Row.

Kenesei, I., R. M. Vago and A. Fenyvesi. 1998. Hungarian. London & New York: Routledge.

Kornai, A. 1994. On Hungarian morphology (Linguistica, Series A: Studia et Dissertationes 14). Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences.

Krallinger, M., F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal and A. Valencia. 2015. CHEMDNER: The drugs and chemical names extraction challenge. Journal of Cheminformatics 7. S1.

Kripke, S. A. 1981. Naming and necessity. Cambridge, MA & Oxford: Blackwell.

Langendonck, W., von. 2007. Theory and typology of proper names. Berlin & New York: Mouton de Gruyter.

LDC/Linguistic Data Consortium LCTL Team. 2006. Simple Named Entity guidelines for less commonly taught languages. Version 6.5.

Lukács, Á. 2001. Szabályok és kivételek: a kettős modell érvényessége a magyarban [Rules and exceptions: The validity of the double model in Hungarian]. In Cs. Pléh and Á. Lukács (eds.) A magyar morfológia pszicholingvisztikája [The psycholinguistics of Hungarian morphology]. Budapest: Osiris Kiadó. 119–152.

Merchant, R., M. E. Okurowski and N. Chinchor. 1996. The Multilingual Entity Task (MET) overview. In Proceedings of the TIPSTER Text Program: Phase II. Vienna, VA: Association for Computational Linguistics, 445–447.

Mill, J. S. 2002. A system of logic. Honolulu: University Press of the Pacific.

Miller, D., R. Schwartz, R. Weischedel and R. Stone. 1999. Named Entity Extraction from broadcast news. In Proceedings of the DARPA Broadcast News Workshop. Herndon, Virginia.

Nadeau, D. and S. Sekine. 2007. A survey of Named Entity Recognition and Classification. Lingvisticae Investigationes 30. 3–26.

Quirk, R. and S. Greenbaum. 1980. A university grammar of English. Harlow: Longman.

Rosch, E. H. 1973. Natural categories. Cognitive Psychology 4. 328–350.

Saussure, F., de. 1959. Course in general linguistics. New York: Philosophical Library.

Simon, E. 2013. Approaches to Hungarian Named Entity Recognition. Doctoral dissertation. Budapest University of Technology and Economics.

Stainton, R. J. (ed.). 2000. Perspectives in the philosophy of language. A concise anthology. Peterborough: Broadview Press.

Sundheim, B. 1995. MUC-6 Named Entity task definition (v2.1). In Proceedings of the Sixth Message Understanding Conference (MUC6).

Szabó, Z. G. 2008. Compositionality. In E. N. Zalta (ed.) The Stanford encyclopedia of philosophy (Winter 2008 edition). Stanford, CA: The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University.

Tjong Kim Sang, E. F. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In D. Roth and A. van den Bosch (eds.) Proceedings of CoNLL-2002. Edmonton: Association for Computational Linguistics. 155–158.

Tjong Kim Sang, E. F. and F. De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In W. Daelemans and M. Osborne (eds.) Proceedings of CoNLL-2003. Edmonton: Association for Computational Linguistics. 142–147.

Zhu, J., V. Uren and E. Motta. 2005. ESpotter: Adaptive named entity recognition for web browsing. In 3rd Conference on Professional Knowledge Management. 518–529.