A study on the worthiness of MWE manually-annotated corpora to train Neural Networks

Emmanuelle Esperança-Rodier

Fiorella Albasini

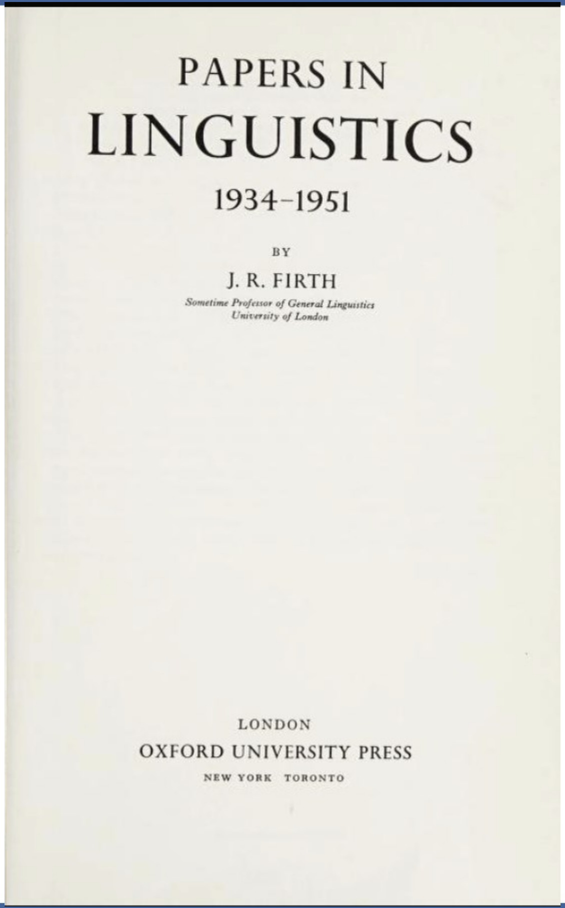
Yacine Haddad

Outline

- Multi Word-Expressions
- MWE annotation
- Experiment
- Inter-annotator agreement
- Issues raised

MWE - Definition

- MultiWord Expressions (MWE) are idiosyncratic expressions made of recurrent word combinations in which the general meaning cannot be understood from the literal meaning of each of its constituents (Firth, 1957)



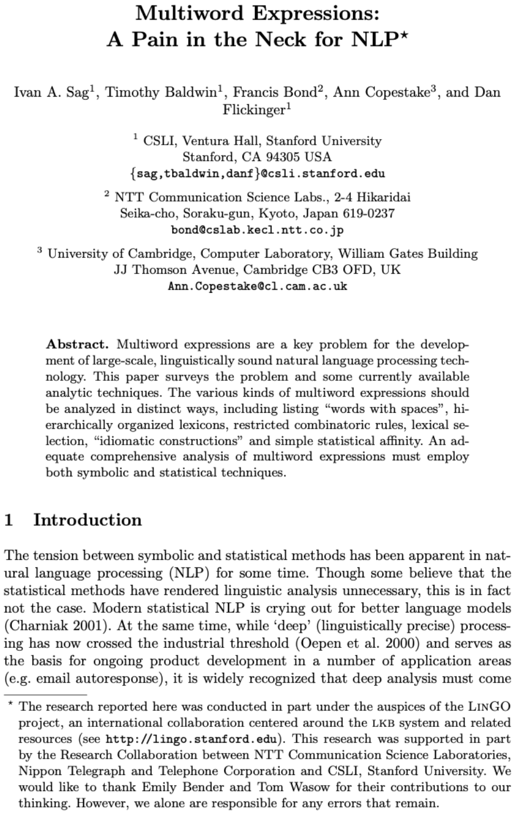
PAPERS IN
LINGUISTICS
1934–1951

BY
J. R. FIRTH
Sometime Professor of General Linguistics
University of London

LONDON
OXFORD UNIVERSITY PRESS
NEW YORK TORONTO

MWE - Definition

- MultiWord Expressions (MWE) are idiosyncratic expressions made of recurrent word combinations in which the general meaning cannot be understood from the literal meaning of each of its constituents (Firth, 1957)
- Sag et al (2002) estimate that their use is equivalent to that of single words in language.



Multiword Expressions: A Pain in the Neck for NLP*

Ivan A. Sag[1], Timothy Baldwin[1], Francis Bond[2], Ann Copestake[3], and Dan Flickinger[1]

[1] CSLI, Ventura Hall, Stanford University
Stanford, CA 94305 USA
{sag,tbaldwin,danf}@csli.stanford.edu

[2] NTT Communication Science Labs., 2-4 Hikaridai
Seika-cho, Soraku-gun, Kyoto, Japan 619-0237
bond@cslab.kecl.ntt.co.jp

[3] University of Cambridge, Computer Laboratory, William Gates Building
JJ Thomson Avenue, Cambridge CB3 OFD, UK
Ann.Copestake@cl.cam.ac.uk

Abstract. Multiword expressions are a key problem for the development of large-scale, linguistically sound natural language processing technology. This paper surveys the problem and some currently available analytic techniques. The various kinds of multiword expressions should be analyzed in distinct ways, including listing "words with spaces", hierarchically organized lexicons, restricted combinatoric rules, lexical selection, "idiomatic constructions" and simple statistical affinity. An adequate comprehensive analysis of multiword expressions must employ both symbolic and statistical techniques.

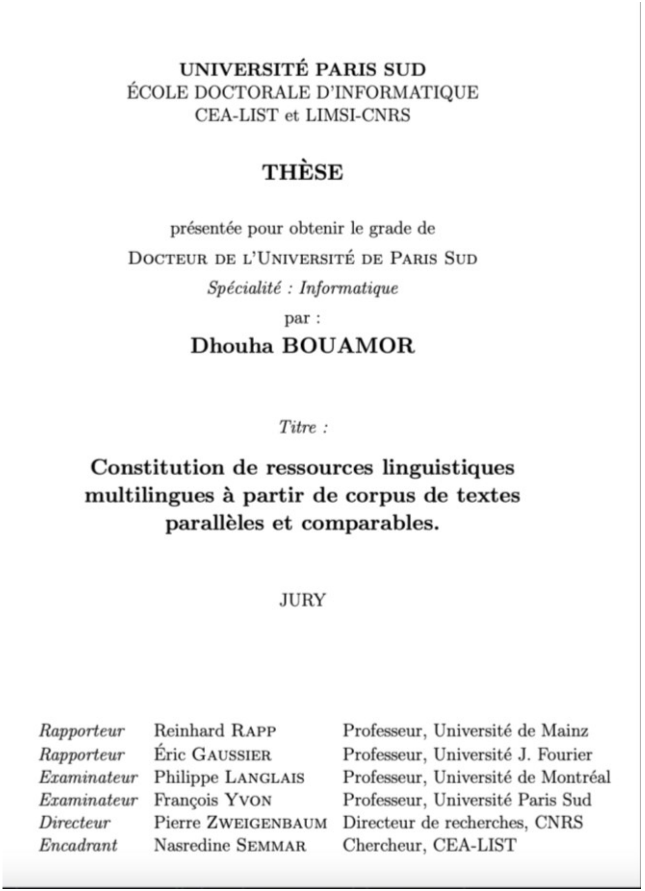
1 Introduction

The tension between symbolic and statistical methods has been apparent in natural language processing (NLP) for some time. Though some believe that the statistical methods have rendered linguistic analysis unnecessary, this is in fact not the case. Modern statistical NLP is crying out for better language models (Charniak 2001). At the same time, while 'deep' (linguistically precise) processing has now crossed the industrial threshold (Oepen et al. 2000) and serves as the basis for ongoing product development in a number of application areas (e.g. email autoresponse), it is widely recognized that deep analysis must come

* The research reported here was conducted in part under the auspices of the LinGO project, an international collaboration centered around the LKB system and related resources (see http://lingo.stanford.edu). This research was supported in part by the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. We would like to thank Emily Bender and Tom Wasow for their contributions to our thinking. However, we alone are responsible for any errors that remain.

MWE – NLP and MT Issues

- MWEs are easily recognized by humans, however, their identification is often problematic in Natural Language Processing (NLP) (Bouamor, 2014).



UNIVERSITÉ PARIS SUD
ÉCOLE DOCTORALE D'INFORMATIQUE
CEA-LIST et LIMSI-CNRS

THÈSE

présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE PARIS SUD

Spécialité : Informatique

par :

Dhouha BOUAMOR

Titre :

Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables.

JURY

Rapporteur	Reinhard RAPP	Professeur, Université de Mainz
Rapporteur	Éric GAUSSIER	Professeur, Université J. Fourier
Examinateur	Philippe LANGLAIS	Professeur, Université de Montréal
Examinateur	François YVON	Professeur, Université Paris Sud
Directeur	Pierre ZWEIGENBAUM	Directeur de recherches, CNRS
Encadrant	Nasredine SEMMAR	Chercheur, CEA-LIST

MWE – NLP and MT Issues

- MWEs are easily recognized by humans, however, their identification is often problematic in Natural Language Processing (NLP) (Bouamor, 2014).
- In machine translation (MT), failing to recognize a MWE is one of the main sources of error. (Constant et al., 2017).

TALN 2011, Montpellier, 27 juin –1er juillet 2011

Intégrer des connaissances linguistiques dans un CRF :
application à l'apprentissage d'un segmenteur-étiqueteur du français

Matthieu Constant[1] Isabelle Tellier[2] Denys Duchier[2]
Yoann Dupont[2] Anthony Sigogne[1] Sylvie Billot[2]
(1) Université Paris-Est, LIGM, CNRS, 5 bd Descartes, Champs-sur-Marne 77454
Marne-la-Vallée cedex 2
(2) LIFO, université d'Orléans, 6 rue Léonard de Vinci
BP 6759, 45067 Orléans cedex 2
mconstan@univ-mlv.fr, isabelle.tellier@univ-orleans.fr,
denys.duchier@univ-orleans.fr, yoann.dupont@etu.univ-orleans.fr,
sigogne@univ-mlv.fr, sylvie.billot@univ-orleans.fr

Résumé. Dans cet article, nous synthétisons les résultats de plusieurs séries d'expériences réalisées à l'aide de CRF (Conditional Random Fields ou "champs markoviens conditionnels") linéaires pour apprendre à annoter des textes français à partir d'exemples, en exploitant diverses ressources linguistiques externes. Ces expériences ont porté sur l'étiquetage morphosyntaxique intégrant l'identification des unités polylexicales. Nous montrons que le modèle des CRF est capable d'intégrer des ressources lexicales riches en unités multi-mots de différentes manières et permet d'atteindre ainsi le meilleur taux de correction d'étiquetage actuel pour le français.

Abstract. In this paper, we synthesize different experiments using a linear CRF (Conditional Random Fields) to annotate French texts from examples, by exploiting external linguistic resources. These experiments especially dealt with part-of-speech tagging including multiword units identification. We show that CRF models allow to integrate, in different ways, large-coverage lexical resources including multiword units and reach state-of-the-art tagging results for French.

Mots-clés : Etiquetage morphosyntaxique, Modèle CRF, Ressources lexicales, Segmentation, Unités polylexicales.

Keywords: Part-of-speech tagging, CRF model, Lexical resources, Segmentation, Multiword units.

MWE – NLP and MT Issues

- MWEs are easily recognized by humans, however, their identification is often problematic in Natural Language Processing (NLP) (Bouamor, 2014).
- In machine translation (MT), failing to recognize a MWE is one of the main sources of error. (Constant et al., 2017).
- Even with the venue of amazing quality results in Neural Machine Translations, NMT still struggle with MWEs (Zaninello and Birch, 2020)

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3816–3825
Marseille, 11–16 May 2020

Multiword Expression aware Neural Machine Translation

Andrea Zaninello[*][†], Alexandra Birch[*]
[*]School of Informatics, University of Edinburgh, United Kingdom
[†]Zanichelli editore, Bologna, Italy
azaninello@zanichelli.it, a.birch@ed.ac.uk

Abstract

Multiword Expressions (MWEs) are a frequently occurring phenomenon found in all natural languages that is of great importance to linguistic theory, natural language processing applications, and machine translation systems. Neural Machine Translation (NMT) architectures do not handle these expressions well and previous studies have rarely addressed MWEs in this framework. In this work, we show that annotation and data augmentation, using external linguistic resources, can improve both translation of MWEs that occur in the source, and the generation of MWEs on the target, and increase performance by up to 5.09 BLEU points on MWE test sets. We also devise a MWE score to specifically assess the quality of MWE translation which agrees with human evaluation. We make available the MWE score implementation – along with MWE-annotated training sets and corpus-based lists of MWEs – for reproduction and extension.

Keywords: multiword expressions, neural machine translation, evaluation

1. Introduction

Multiword Expressions (MWEs) are a pervasive phenomenon in all natural languages to the point that, according to some studies, they represent approximately half of a language's lexicon (Jackendoff, 1995). They also challenge NLP applications because of their often unpredictable morpho-syntactic and lexico-semantic behaviour (Villavicencio et al., 2005). We call a MWE an expression that is composed of two or more words working as a unit with respect to some levels of linguistic analysis (Calzolari et al., 2002); a MWE displays idiosyncratic properties that cannot be explained solely on the basis of regular syntactic and semantic rules (Everaert et al., 2014) and is generally characterised by some degree of conventionality (Baldwin and Kim, 2010; Constant et al., 2017).

In the last few years, Neural Machine Translation (NMT) has proved the best performing framework compared to previous methodologies, with neural architectures producing ever more natural-sounding target language. Even so, NMT output is sometimes a poor translation of the source sentence (Nguyen and Chiang, 2018) and it is therefore important to investigate specific linguistic phenomena and improve translation quality not only in terms of standard measurements.

Previously dominant phrase-based and syntax-based Statistical Machine Translation (SMT) techniques (Koehn et al., 2007; Junczys-Dowmunt et al., 2016) naturally take into account phrasal components, and there has been significant research on MWEs in these frameworks; however, for NMT, due to a lack of phrasal segmentation, it is less obvious how to address specific language phenomena such as MWEs. Moreover, while standard metrics are effective in terms of system comparison, their ability to account for more fine-grained improvements in MT is less straightforward (Callison-Burch et al., 2006), and their effectiveness has been questioned. Therefore, evaluating the performance of NMT architectures in translating MWEs remains an open challenge.

The aim of this study is to empirically verify whether integrating information on MWEs either through targeted training examples or through explicit annotation in the target language can help disambiguating between simple phrasal units and non-compositional expressions, and thus be beneficial to NMT. In our first approach, we try augmenting our training data with entries from a bilingual and a monolingual MWE dictionary, adding a relatively small number of instances (10% and 2% of the original data, respectively), both in isolation and in their sentence context from usage examples provided. The second approach takes a MWE annotation tool, and labels MWEs on the source. We either concatenate MWE into one word or we use factors to indicate if they form part of a MWE.

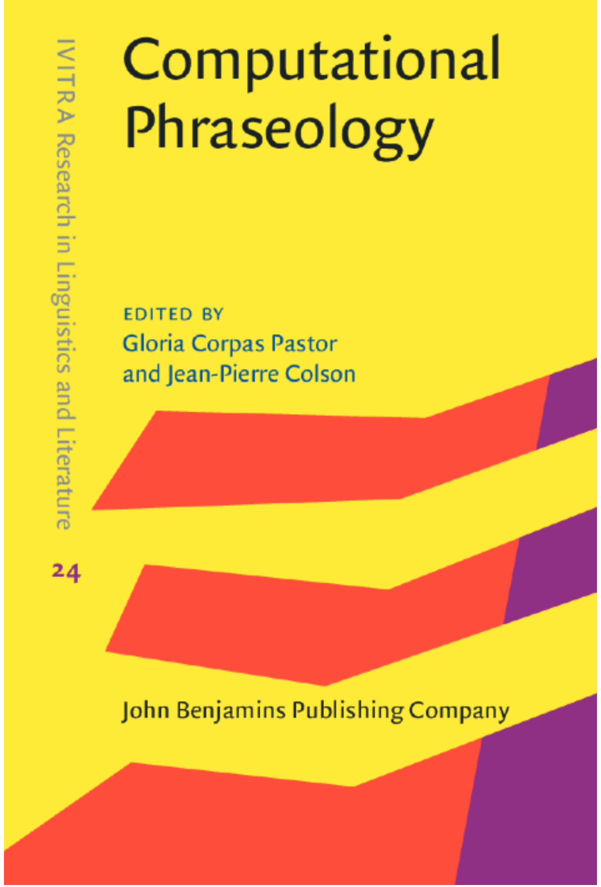
We show that for a test set comprised of genuinely non-compositional MWEs the NMT output is of extremely low quality, indicating that these models struggle to handle these examples, especially in the small training data condition. We also show that all our methods improve translation in general and MWE translation in particular. The method of including MWE in context, with backtranslation to recreate the source side, does well in the low resource setting, but given the small number of genuine examples is not scalable. Our approach of labelling MWEs does however extend to improving translation in a large resource experiment.

In order to further analyse our results, we propose a novel evaluation metric (the *Score_mwe*) that specifically evaluates how well MWEs on the source side are translated. It needs a test set with human annotated MWEs on the source and their translation in the reference. It uses the Levenshtein distance to find the closest matching word in the hypothesis and rewards partial matches at the character level. We compare our novel metric with manual evaluation and show that it agrees with human judgments.

In this paper we limit our study to one language pair (from English to Italian) and to one specific neural architecture, but our methods can easily be extended to other language combinations or different NMT frameworks. We also rely on human curated resources in order to prove their value to NMT, and in future work we plan to consider automatically extracted MWE lexicons and unsupervised taggers.

MWE – NLP and MT Issues

- MWEs are easily recognized by humans, however, their identification is often problematic in Natural Language Processing (NLP) (Bouamor, 2014).
- In machine translation (MT), failing to recognize a MWE is one of the main sources of error. (Constant et al., 2017).
- Even with the venue of amazing quality results in Neural Machine Translations, NMT still struggle with MWEs (Zaninello and Birch, 2020)
- Colson (2020) reports that Google Translate made mistakes in about 40% of MWE translations.



IVITRA Research in Linguistics and Literature

24

Computational Phraseology

EDITED BY
Gloria Corpas Pastor
and Jean-Pierre Colson

John Benjamins Publishing Company

MWE – NLP and MT Issues

- MWEs are easily recognized by humans, however, their identification is often problematic in Natural Language Processing (NLP) (Bouamor, 2014).
- In machine translation (MT), failing to recognize a MWE is one of the main sources of error. (Constant et al., 2017).
- Even with the venue of amazing quality results in Neural Machine Translations, they still struggle with MWEs (Zaninello and Birch, 2020)
- Colson (2020) reports that Google Translate made mistakes in about 40% of MWE translations.

Identifying MWE

Mind-The-Gap

mwetoolkit^3

PARSEME.FR

Existing MWE annotated Corpora

- Laporte et al. (2008a; 2008b)

A French Corpus Annotated for Multiword Nouns

Éric Laporte, Takuya Nakamura, Stavroula Voyatzi
Université Paris-E
IGM-Labinfo
5, Boulevard Descartes, Cham
77454 Marne-la-Vallée Cede

E-mail: eric.laporte@univ-paris-est.fr, nakamura@

Abstract

This paper presents a French corpus annotated for multiword nouns. This cor
extraction, as well as in deep and shallow syntactic parsing. We delimit wh
task; we describe the resources and methods we used for the annotation; and
available at *http://infolingu.univ-mlv.fr/* under the LGPLLR license.

1. Introduction

Recognizing multiword nouns such as *groupes de pression* 'lobbies' in texts is useful for information retrieval and extraction because of the information that such nouns can convey. In particular, in specialized languages, most of the technical and terminological information is concentrated in multiword nouns. In addition, such recognition is likely to help resolving prepositional attachment during shallow or deep parsing: some multiword nouns contain internal prepositional phrases, and in many cases, recognising them rules out analyses where they are complements of verbs, adjectives or other nouns (Blanc *et al.*, 2007). In the case of English, the same is true for the analysis of noun sequences (Vadas & Curran, 2007).

The quality of the recognition of multiword nouns depends on algorithms, but also on resources. We created a corpus of French texts annotated with multiword nouns. This corpus is freely available on the web with LGPLLR license. In this article, we survey related work, we define the target of our annotation effort, we describe the method implemented and we analyse the corpus obtained.

2. Related work

Many problems related with the notion of multiword expression (MWE) in general have been studied by linguists and lexicologists (e.g. Downing, 1977; Sag *et al.*, 2001; Girju, 2005; as regards French multiword nouns: Silberztein, 1993), but textual resources annotated for MWEs are still rare and small. In the Grace corpus (Rajman *et al.*, 1997), most MWEs are ignored. In the French Treebank (Abeillé *et al.*, 2003), multiword nouns are annotated as such. We are not aware of other available French corpora annotated with multiword nouns. In other languages, including English, corpora annotated with MWEs are rare and small as well. In the Penn Treebank (Marcus *et al.*, 1993), even such frozen nouns as *stock market* are not annotated as MWEs. Subirats & Sato (2004) report an experiment of annotating MWUs, including multiword nouns, in a Spanish corpus, and Mota *et al.* (2004) and Ranchhod (2005) in a Portuguese corpus, but

A French Corpus Annotated for Multiword Expressions with Adverbial Function

Eric Laporte, Takuya Nakamura, Stavroula Voyatzi
Université Paris-Est
Institut Gaspard-Monge - LabInfo
5, Boulevard Descartes, Champs-sur-Marne
77454 Marne-la-Vallée Cedex 2 (France)
E-mail: eric.laporte@univ-paris-est.fr, nakamura@univ-mlv.fr, voyatzi@univ-mlv.fr

Abstract

This paper presents a French corpus annotated for multiword expressions (MWEs) with adverbial function. This corpus is designed for investigation on information retrieval and extraction, as well as on deep and shallow syntactic parsing. We delimit which kind of MWEs we annotated, we describe the resources and methods we used for the annotation, and we briefly comment the results. The annotated corpus is available at *http://infolingu.univ-mlv.fr/* under the LGPLLR license.

1. Introduction

Recognising multiword adverbs such as *à long terme* 'in the long run' in texts is likely to be useful for information retrieval and extraction because of the information that such adverbials can convey. In addition, it is likely to help resolving prepositional attachment during shallow or deep parsing: most multiword adverbs have the superficial syntax of prepositional phrases; in many cases, recognising them rules out analyses where they are arguments or noun modifiers.

The quality of the recognition of multiword adverbs depends on algorithms, but also on resources. We created a corpus of French texts annotated with multiword adverbs. In this article, we survey related work, we define the target of our annotation effort, we describe the method we have implemented and we analyse the corpus obtained. This corpus will be made freely available on the web under the LGPLLR license when this article is published.

2. Related work

Corpora annotated with multiword adverbs are rare and small[1]. In the Grace corpus (Rajman *et al.*, 1997), most multiword units are ignored. In the French Treebank (Abeillé *et al.*, 2003), prepositional phrases and adverbs are annotated with a binary feature ('compound') which indicates whether they are multiword units; the distinction between whether prepositional phrases are verb modifiers, noun modifiers or objects appears only in the function-annotated part of the Treebank (350 000 words). We are not aware of other available French corpora annotated with multiword adverbs. In other languages, including English, corpora annotated with multiword units are rare and small as well.

3. Target of annotation

The target of our annotation effort is defined by the intersection of two criteria: (i) multiword expressions and (ii) adverbial function. In this section, we define both criteria in more detail, we define the features that we included in the annotations, and we describe the corpus.

3.1 Multiword expression criterion

For this work, we considered a phrase composed of several words to be a multiword expression if some or all of their elements are frozen together in the sense of (Gross, 1986), that is, if their combination does not obey productive rules of syntactic and semantic compositionality. In the following example, *de nos jours* ('nowadays', lit. 'of our days') is a multiword adverb:

(1) *Il est facile* ***de nos jours*** *de s'informer*
'It is easy to get informed **nowadays**'

This criterion ensures a complementarity between lexicon and grammar. In other words, it tends to ensure[2] that any combination of linguistic elements which is licit in the language, but is not represented in syntactic-semantic grammars, will be stored in lexicons.

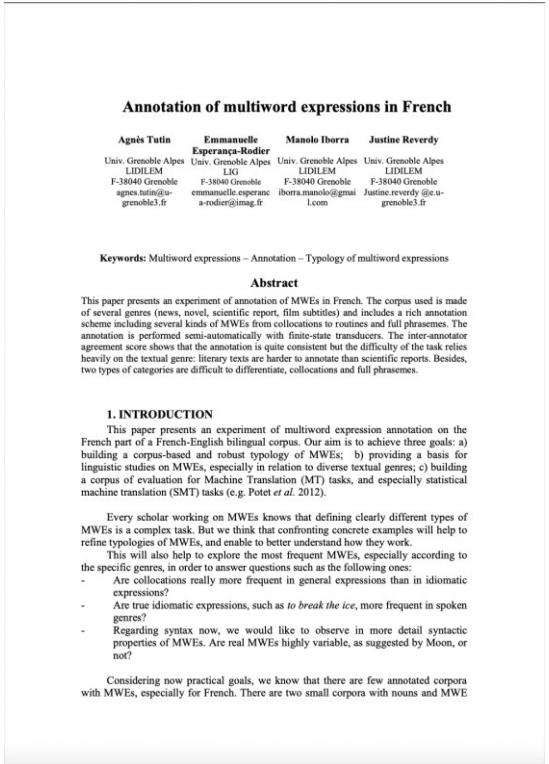
Syntactic-semantic compositionality is usually defined as follows (Freckleton, 1985; Machonis, 1985; Silberztein, 1993; Lamiroy, 2003): a combination of linguistic elements is compositional if and only if its meaning can be computed from its elements. This is also our conception. However, in this definition, we consider that the possibility of computing the meaning of phrases from their elements is of any interest only if it is a better solution than storing the same phrases in lexicons, i.e. if

[1] Several reasons explain this lack of interest. Firstly, adverbials are usually felt as less useful than nouns for information retrieval and extraction. Secondly, many multiword adverbs are difficult to distinguish from prepositional phrases assuming other syntactic functions, such as arguments or noun modifiers: the distinction is hardly correlated to any material markers in texts and lies in complex linguistic notions (Villavicencio, 2002; Merlo, 2003). The task is therefore felt as too difficult by most researchers in language processing, whose main background is in information technology. However, the distinction in question is essential to identifying the semantic core of a sentence, and the availability of a larger corpus of annotated text is likely to shed light on the problems posed by this task.

[2] That can be empirically checked only after a lexicon and a grammar for the same language are complete and compatible.

Existing MWE annotated Corpora

- Laporte et al. (2008a; 2008b)
- PolyCorp, Tutin (2016) & Tutin and Esperança-Rodier (2019)



Annotation of multiword expressions in French

Agnès Tutin
Univ. Grenoble Alpes
LIDILEM
F-38040 Grenoble
agnes.tutin@u-grenoble3.fr

Emmanuelle Esperança-Rodier
Univ. Grenoble Alpes
LIG
F-38040 Grenoble
emmanuelle.esperanca-rodier@imag.fr

Manolo Iborra
Univ. Grenoble Alpes
LIDILEM
F-38040 Grenoble
iborra.manolo@gmail.com

Justine Reverdy
Univ. Grenoble Alpes
LIDILEM
F-38040 Grenoble
Justine.reverdy @e.u-grenoble3.fr

Keywords: Multiword expressions – Annotation – Typology of multiword expressions

Abstract

This paper presents an experiment of annotation of MWEs in French. The corpus used is made of several genres (news, novel, scientific report, film subtitles) and includes a rich annotation scheme including several kinds of MWEs from collocations to routines and full phrasemes. The annotation is performed semi-automatically with finite-state transducers. The inter-annotator agreement score shows that the annotation is quite consistent but the difficulty of the task relies heavily on the textual genre: literary texts are harder to annotate than scientific reports. Besides, two types of categories are difficult to differentiate, collocations and full phrasemes.

1. INTRODUCTION

This paper presents an experiment of multiword expression annotation on the French part of a French-English bilingual corpus. Our aim is to achieve three goals: a) building a corpus-based and robust typology of MWEs; b) providing a basis for linguistic studies on MWEs, especially in relation to diverse textual genres; c) building a corpus of evaluation for Machine Translation (MT) tasks, and especially statistical machine translation (SMT) tasks (e.g. Potet *et al.* 2012).

Every scholar working on MWEs knows that defining clearly different types of MWEs is a complex task. But we think that confronting concrete examples will help to refine typologies of MWEs, and enable to better understand how they work.

This will also help to explore the most frequent MWEs, especially according to the specific genres, in order to answer questions such as the following ones:

- Are collocations really more frequent in general expressions than in idiomatic expressions?
- Are true idiomatic expressions, such as *to break the ice*, more frequent in spoken genres?
- Regarding syntax now, we would like to observe in more detail syntactic properties of MWEs. Are real MWEs highly variable, as suggested by Moon, or not?

Considering now practical goals, we know that there are few annotated corpora with MWEs, especially for French. There are two small corpora with nouns and MWE

Existing MWE annotated Corpora

- Laporte et al. (2008a; 2008b)
- PolyCorp, Tutin (2016) & Tutin and Esperança-Rodier (2019)
- SzegedParalellFX English–Hungarian, Vincze (2012)

Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus

Veronika Vincze

Hungarian Academy of Sciences
Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

Abstract

In this paper, we describe the first English–Hungarian parallel corpus annotated for light verb constructions, which contains 14,261 sentence alignment units. Annotation principles and statistical data on the corpus are also provided, and English and Hungarian data are contrasted. On the basis of corpus data, a database containing pairs of English–Hungarian light verb constructions has been created as well. The corpus and the database can contribute to the automatic detection of light verb constructions and they can enhance performance in several fields of NLP (e.g. parsing, information extraction/retrieval and machine translation).

Keywords: light verb constructions, English–Hungarian parallel corpus, multilinguality

1. Introduction

In natural language processing (NLP), one of the most challenging tasks is the proper treatment of multiword expressions (MWEs). MWEs are lexical items that can be decomposed into single words and display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Sag et al., 2002; Calzolari et al., 2002). Light verb constructions form a subtype of multiword expressions. They consist of a nominal and a verbal component where the noun is usually taken in one of its literal senses but the verb loses its original sense to some extent, e.g. *to give advice*, *to take into account*, *the problem lies (in)*. They are frequent in language use and because of their idiosyncratic behavior, they often pose a problem to NLP systems.

In this paper, we describe SzegedParalellFX, the first English–Hungarian parallel corpus annotated for light verb constructions. We believe that the corpus can contribute to the research on multiword expressions and more specifically, to the development of algorithms aiming at detecting light verb constructions.

The structure of the paper is as follows. First, related corpora and related work on the NLP treatment of multiword expressions are presented. Then the corpus is described together with annotation principles. Some statistical data on corpus data are also provided, which is followed by a qualitative analysis and a comparison of English and Hungarian data. The paper concludes with illustrating how the corpus and the database can be exploited in several fields of NLP.

2. Related work

Lately, multiword expressions have been received special interest in the NLP research community (Rayson et al., 2010). This also holds for multiword verbs since they constitute a subtype of multiword expressions, e.g. Sag et al. (2002) classify them as a subtype of lexicalized phrases and flexible expressions. The automatic identification of multiword verbs has been studied in several languages. Cook et al. (2007) differentiate between literal and idiomatic usages of verb and noun constructions in English. Their basic hypothesis is that the canonical form of each construction occurs mostly in idioms since they show syntactic variation to a lesser degree than constructions in literal usage. Hence, they make use of syntactic fixedness of idioms when developing their unsupervised method.

Van de Cruys and Moirón (2007) describe a semantic-based method for identifying verb-preposition-noun combinations in Dutch. Their method relies on selectional preferences for both the noun and the verb and they also make use of automatic noun clustering when considering the selection of semantic classes of nouns for each verb.

Bannard (2007) seeks to identify verb and noun constructions in English on the basis of syntactic fixedness. He examines whether the noun can have a determiner or not, whether the noun can be modified and whether the construction can have a passive form, which features are exploited in the identification of the constructions.

Gurrutxaga and Alegria (2011) extract idioms and light verb constructions from Basque texts by employing statistical methods. Since Basque is a free word-order language, they hypothesized that a wider window would yield more significant cooccurrence statistics, however, their initial experiments did not confirm this.

Tu and Roth (2011) classify verb + noun object pairs as being light verb constructions or not. They operate with both contextual and statistical features and conclude that on ambiguous examples, local contextual features perform better.

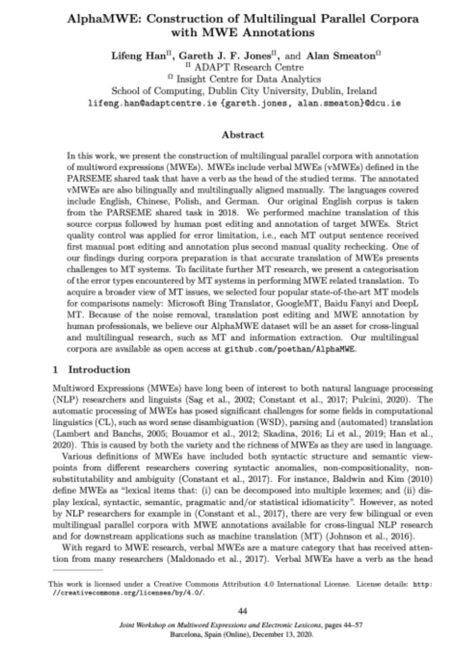
Vincze et al. (2011a) exploit shallow morphological features in identifying English light verb constructions and the domain specificity of the problem is emphasized in Nagy T. et al. (2011).

Parallel corpora are of high importance in the automatic identification of multiword expressions: it is usually one-to-many correspondence that is exploited when designing methods for detecting multiword expressions. On the other hand, aligned parallel corpora can also enhance the identification of multiword expressions in different languages: if an algorithm is implemented for one language, data from the other language can also be gathered with the help of aligned units.

2381

Existing MWE annotated Corpora

- Laporte et al. (2008a; 2008b)
- PolyCorp, Tutin (2016) & Tutin and Esperança-Rodier (2019)
- SzegedParalellFX English–Hungarian, Vincze (2012)
- AlphaMWE, Han et al. (2020)



AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations

Lifeng Han[Π], **Gareth J. F. Jones**[Π], and **Alan Smeaton**[Ω]
[Π] ADAPT Research Centre
[Ω] Insight Centre for Data Analytics
School of Computing, Dublin City University, Dublin, Ireland
lifeng.han@adaptcentre.ie {gareth.jones, alan.smeaton}@dcu.ie

Abstract

In this work, we present the construction of multilingual parallel corpora with annotation of multiword expressions (MWEs). MWEs include verbal MWEs (vMWEs) defined in the PARSEME shared task that have a verb as the head of the studied terms. The annotated vMWEs are also bilingually and multilingually aligned manually. The languages covered include English, Chinese, Polish, and German. Our original English corpus is taken from the PARSEME shared task in 2018. We performed machine translation of this source corpus followed by human post editing and annotation of target MWEs. Strict quality control was applied for error limitation, i.e., each MT output sentence received first manual post editing and annotation plus second manual quality rechecking. One of our findings during corpora preparation is that accurate translation of MWEs presents challenges to MT systems. To facilitate further MT research, we present a categorisation of the error types encountered by MT systems in performing MWE related translation. To acquire a broader view of MT issues, we selected four popular state-of-the-art MT models for comparisons namely: Microsoft Bing Translator, GoogleMT, Baidu Fanyi and DeepL MT. Because of the noise removal, translation post editing and MWE annotation by human professionals, we believe our AlphaMWE dataset will be an asset for cross-lingual and multilingual research, such as MT and information extraction. Our multilingual corpora are available as open access at github.com/poethan/AlphaMWE.

1 Introduction

Multiword Expressions (MWEs) have long been of interest to both natural language processing (NLP) researchers and linguists (Sag et al., 2002; Constant et al., 2017; Pulcini, 2020). The automatic processing of MWEs has posed significant challenges for some fields in computational linguistics (CL), such as word sense disambiguation (WSD), parsing and (automated) translation (Lambert and Banchs, 2005; Bouamor et al., 2012; Skadina, 2016; Li et al., 2019; Han et al., 2020). This is caused by both the variety and the richness of MWEs as they are used in language.

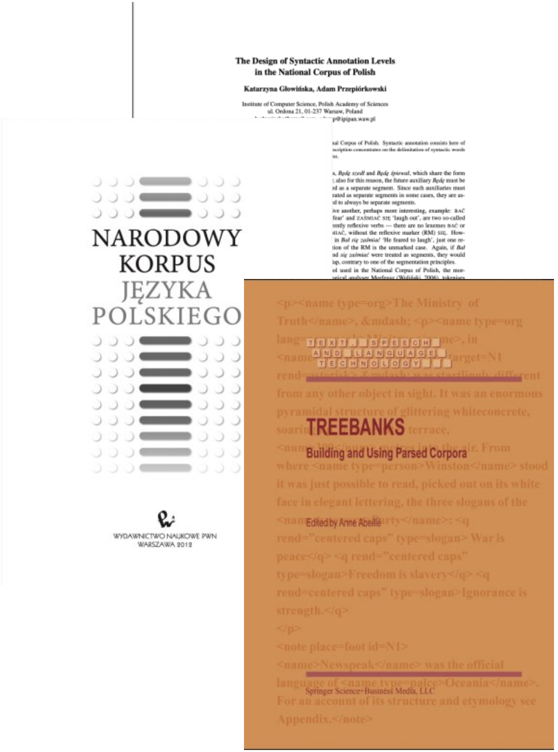
Various definitions of MWEs have included both syntactic structure and semantic viewpoints from different researchers covering syntactic anomalies, non-compositionality, non-substitutability and ambiguity (Constant et al., 2017). For instance, Baldwin and Kim (2010) define MWEs as "lexical items that: (i) can be decomposed into multiple lexemes; and (ii) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". However, as noted by NLP researchers for example in (Constant et al., 2017), there are very few bilingual or even multilingual parallel corpora with MWE annotations available for cross-lingual NLP research and for downstream applications such as machine translation (MT) (Johnson et al., 2016).

With regard to MWE research, verbal MWEs are a mature category that has received attention from many researchers (Maldonado et al., 2017). Verbal MWEs have a verb as the head

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 44–57
Barcelona, Spain (Online), December 13, 2020.

Existing MWE annotated Corpora

- Laporte et al. (2008a; 2008b)
- PolyCorp, Tutin (2016) & Tutin and Esperança-Rodier (2019)
- SzegedParalellFX English–Hungarian, Vincze (2012)
- AlphaMWE, Han et al. (2020)
- Treebanks: Abeillé et al. (2003)
  Głowińska & Przepiórkowski (2010)
  Głowińska (2012)



The Design of Syntactic Annotation Levels
in the National Corpus of Polish

Katarzyna Głowińska, Adam Przepiórkowski

Institute of Computer Science, Polish Academy of Sciences
ul. Ordona 21, 01-237 Warsaw, Poland

NARODOWY KORPUS JĘZYKA POLSKIEGO

WYDAWNICTWO NAUKOWE PWN
WARSZAWA 2012

TEXT, SPEECH AND LANGUAGE TECHNOLOGY

TREEBANKS
Building and Using Parsed Corpora

Edited by Anne Abeillé

Springer Science+Business Media, LLC

Existing MWE annotated Corpora

- Laporte et al. (2008a; 2008b)
- PolyCorp, Tutin (2016) & Tutin and Esperança-Rodier (2019)
- SzegedParalellFX English–Hungarian, Vincze (2012)
- AlphaMWE, Han et al. (2020)
- Treebanks: Abeillé et al. (2003)
  Głowińska & Przepiórkowski (2010)
  Głowińska (2012)

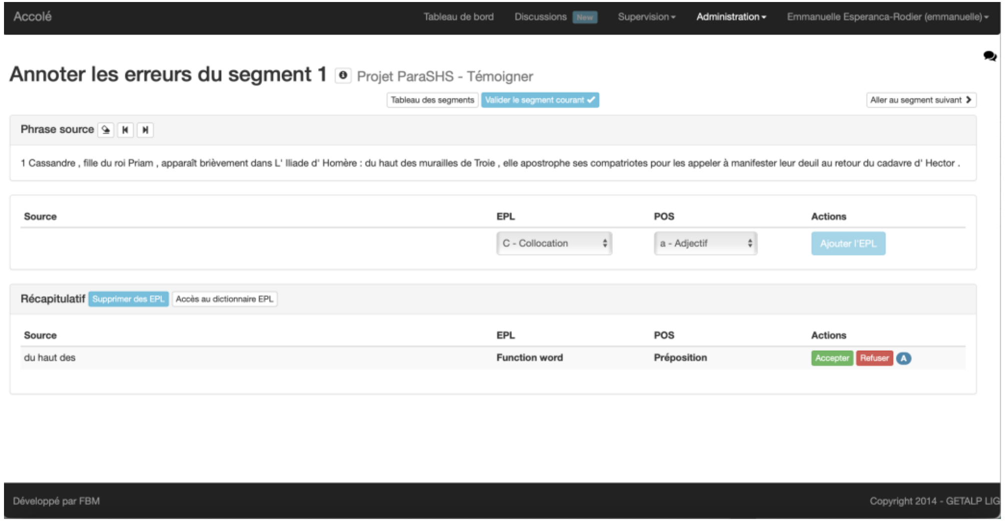
Our concerns

- Figure out if the annotation made by human annotators could provide high quality corpora in a reasonable quantity
- Is the quality of the human annotations consistent among the different annotators?
- Is the size of our corpus big enough for NN systems?
- Focus on the inter-annotator agreement
- Annotate a French corpus

Protocol

ACCOLÉ
(Esperança-Rodier and Brunet-Manquat, 2019)
with TYPOLOGY (Tutin, 2016)

ParaSHS-Témoigner
(Kraif, 2018)



Annoter les erreurs du segment 1 Projet ParaSHS - Témoigner

Phrase source

1 Cassandre , fille du roi Priam , apparaît brièvement dans L' Iliade d' Homère : du haut des murailles de Troie , elle apostrophe ses compatriotes pour les appeler à manifester leur deuil au retour du cadavre d' Hector .

Source	EPL	POS	Actions
	C - Collocation	a - Adjectif	Ajouter l'EPL

Récapitulatif

Source	EPL	POS	Actions
du haut des	Function word	Préposition	Accepter Refuser A

ParaSHS-Témoigner
MWE ANNOTATED
(Kraif, 2018)

3 356
annotated
MWEs

Multiword expressions		Examples
Idioms	frozen multiword expressions	cul de sac (fr)/ dead end; prendre en compte (fr)/ take into account
Collocations	preferred binary association, including light verb constructions	gros fumeur (fr)/ heavy smoker; faire une promenade (fr)/ to take a walk
Functional Multiword Expressions	functional adverbs, prepositions, conjunctions, determiners, pronouns.	c’est pourquoi (fr)/ that is why; d’autre part (fr)/ on the other hand; insofar as
Pragmatic MWEs	multiword expressions related to specific speech situations.	de rien (fr)/ You’re welcome; à plus tard (fr)/ see you later.
Proverbs		Pierre qui roule n’amasse pas mousse (fr)/ A rolling stone gathers no moss
Complex terms		natural language processing
Multiword Named entities		Université Grenoble Alpes; the European Union;
Routine formulae	routines generally associated to rhetorical functions	force est de constater (fr)/ it must be noted.

Agnès Tutin, Emmanuelle Esperança-Rodier, Manolo Iborra, Justine Reverdy. Annotation of multiword expressions in French. *European Society of Phraseology Conference (EUROPHRAS 2015)*, Jun 2015, Malaga, Spain. pp.60-67.

Inter-annotator agreement - Methodology

- Metric given during the SemEval'13 (International Workshop on Semantic Evaluation) adapted to MWE annotation
  - no gold standard
  - use one of the annotators as the gold standard (gold annotator)
  - compare the gold annotator annotations with the ones from the other annotators, two by two.

Inter-annotator agreement - Methodology

- 4 cases to measure the precision, recall and F-measure between the annotators:
  - Strict evaluation (exact-boundary and type matching).
  - Exact boundary matching (regardless to the type).
  - Partial boundary matching (regardless to the type).
  - Type matching (some overlap between the annotated output and the golden standard is required).

Inter-annotator agreement - Methodology

- 4 cases relate to the 5 MUC (Message Understanding Conference) axis:
  - Correct (COR): annotator output DOES correspond to gold annotator
  - Incorrect (INC): annotator output does NOT correspond to gold annotator
  - Partial (PAR): annotator output and gold annotator are somehow similar but not identical
  - Missing (MIS): Gold annotator annotation not captured by the annotator
  - Spurius (SPU): annotator output not present in the gold annotator annotation

[…] elle rappelle les crimes enfouis à l'origine de la malédiction des Atrides qu'actualisent **une nouvelle fois** l'assassinat d'Agamemnon par Clytemnestre et le matricide commis par Oreste.

Gold Annotator		**Annotator**		**Evaluation Scheme**			
Phrase	**MWE Type**	**Phrase**	**MWE Type**	**Type**	**Partial**	**Exact**	**Strict**
		une nouvelle fois	**Collocation**	**SPU**	**SPU**	**SPU**	**SPU**
Agamemnon d’Eschyle	Named Entity	*l’Agamemnon d’Eschyle*	Named Entity	COR	PAR	INC	INC
au bas de	Function Word	*au bas de*	Collocation	INC	COR	COR	INC
manifester leur deuil	Collocation	*manifester leur deuil*	Collocation	COR	COR	COR	COR
Les droits de l’homme	Full Phraseme	*droits de l’homme*	Collocation	INC	PAR	INC	INC

Prophétesse inspirée par Apollon (à partir de l'**Agamemnon d'Eschyle**) ou faisant bon usage de sa raison (dans nombre de versions modernes), elle devient une figure [...]

Gold Annotator		Annotator		Evaluation Scheme			
Phrase	**MWE Type**	**Phrase**	**MWE Type**	**Type**	**Partial**	**Exact**	**Strict**
		une nouvelle fois	Collocation	SPU	SPU	SPU	SPU
Agamemnon d'Eschyle	Named Entity	*l'Agamemnon d'Eschyle*	Named Entity	COR	PAR	INC	INC
au bas de	Function Word	*au bas de*	Collocation	INC	COR	COR	INC
manifester leur deuil	Collocation	*manifester leur deuil*	Collocation	COR	COR	COR	COR
Les droits de l'homme	Full Phraseme	*droits de l'homme*	Collocation	INC	PAR	INC	INC

[...] il a proféré le terrible constat – non de la disparition des témoins, lui qui allait le 11 avril 1987 se jeter du troisième étage **au bas de** l'escalier de son immeuble.

Gold Annotator		Annotator		Evaluation Scheme			
Phrase	**MWE Type**	**Phrase**	**MWE Type**	**Type**	**Partial**	**Exact**	**Strict**
		une nouvelle fois	Collocation	SPU	SPU	SPU	SPU
Agamemnon d'Eschyle	Named Entity	*l'Agamemnon d'Eschyle*	Named Entity	COR	PAR	INC	INC
au bas de	**Function Word**	***au bas de***	**Collocation**	**INC**	**COR**	**COR**	**INC**
manifester leur deuil	Collocation	*manifester leur deuil*	Collocation	COR	COR	COR	COR
Les droits de l'homme	Full Phraseme	*droits de l'homme*	Collocation	INC	PAR	INC	INC

Cassandre, fille du roi Priam, apparaît brièvement dans L'Iliade d'Homère : du haut des murailles de Troie, elle apostrophe ses compatriotes pour les appeler à **manifester leur deuil** au retour du cadavre d' Hector.

Gold Annotator		Annotator		Evaluation Scheme			
Phrase	**MWE Type**	**Phrase**	**MWE Type**	**Type**	**Partial**	**Exact**	**Strict**
		une nouvelle fois	Collocation	SPU	SPU	SPU	SPU
Agamemnon d'Eschyle	Named Entity	*l'Agamemnon d'Eschyle*	Named Entity	COR	PAR	INC	INC
au bas de	Function Word	*au bas de*	Collocation	INC	COR	COR	INC
manifester leur deuil	**Collocation**	***manifester leur deuil***	**Collocation**	**COR**	**COR**	**COR**	**COR**
Les droits de l'homme	Full Phraseme	*droits de l'homme*	Collocation	INC	PAR	INC	INC

[…] en considérant qu'un questionnement sur les fondements du monde que nous voulons, résolument ancré sur **les droits de l'homme**, doit passer par Auschwitz, tout autant que par la critique de modèles […]

Gold Annotator		**Annotator**		**Evaluation Scheme**			
Phrase	**MWE Type**	**Phrase**	**MWE Type**	**Type**	**Partial**	**Exact**	**Strict**
		une nouvelle fois	Collocation	SPU	SPU	SPU	SPU
Agamemnon d'Eschyle	Named Entity	*l'Agamemnon d'Eschyle*	Named Entity	COR	PAR	INC	INC
au bas de	Function Word	*au bas de*	Collocation	INC	COR	COR	INC
manifester leur deuil	Collocation	*manifester leur deuil*	Collocation	COR	COR	COR	COR
Les droits de l'homme	**Full Phraseme**	***droits de l'homme***	**Collocation**	**INC**	**PAR**	**INC**	**INC**

Inter-annotator agreement - Examples

Gold Annotator		Annotator		Evaluation Scheme			
Phrase	**MWE Type**	**Phrase**	**MWE Type**	**Type**	**Partial**	**Exact**	**Strict**
		une nouvelle fois	Collocation	SPU	SPU	SPU	SPU
Agamemnon d'Eschyle	Named Entity	*l'Agamemnon d'Eschyle*	Named Entity	COR	PAR	INC	INC
au bas de	Function Word	*au bas de*	Collocation	INC	COR	COR	INC
manifester leur deuil	Collocation	*manifester leur deuil*	Collocation	COR	COR	COR	COR
Les droits de l'homme	Full Phraseme	*droits de l'homme*	Collocation	INC	PAR	INC	INC

Inter-annotator agreement - Metrics

- 2 values to be calculated:
  - « possible (POS) » sum of annotations of gold annotator (true positive + false negative) for each of the 4 cases:
    - POSSIBLE(POS) = COR+INC+PAR+MIS=TP+FN
  - « actual (ACT) » sum of the effective annotations of annotator (true positive + false positive) for each of the 4 cases
    - ACTUAL(ACT) = COR+INC+PAR+SPU=TP+FP

Inter-annotator agreement - Metrics

- Standard precision and Standard recall for Exact Cases

$$\text{Precision}_{\text{Std}} = \frac{COR}{ACT} = \frac{TP}{TP+FP}$$

$$\text{Recall}_{\text{Std}} = \frac{COR}{POS} = \frac{TP}{TP+FN}$$

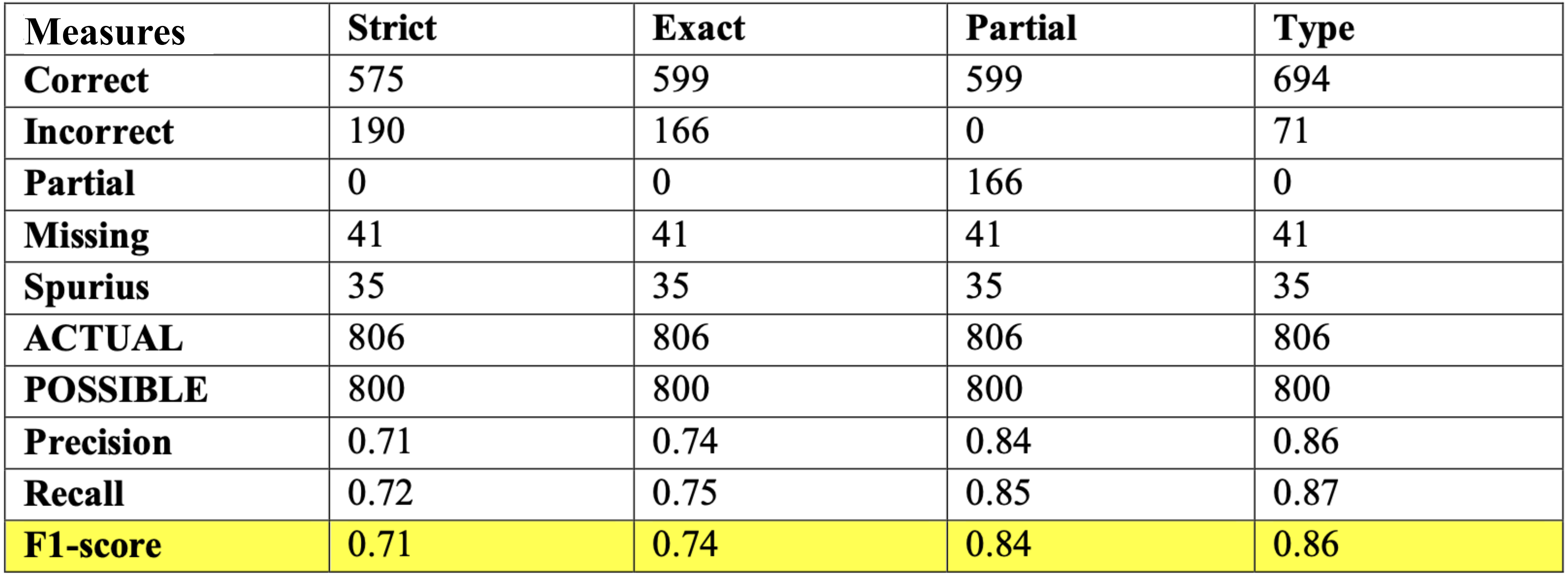
Inter-annotator agreement - Metrics

- Precision + Partial Case and Recall + Partial Case

$$\text{Precision}_{\text{PC}} = \frac{COR+0.5\times PAR}{ACT} = \frac{TP+0.5\times PAR}{TP+FP}$$

$$\text{Recall}_{\text{PC}} = \frac{COR+0.5\times PAR}{POS} = \frac{TP+0.5\times PAR}{TP+FN}$$

Inter-annotator agreement - Results



Measures	Strict	Exact	Partial	Type
Correct	575	599	599	694
Incorrect	190	166	0	71
Partial	0	0	166	0
Missing	41	41	41	41
Spurius	35	35	35	35
ACTUAL	806	806	806	806
POSSIBLE	800	800	800	800
Precision	0.71	0.74	0.84	0.86
Recall	0.72	0.75	0.85	0.87
F1-score	0.71	0.74	0.84	0.86

Conclusion

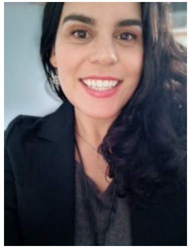
- Human annotation is:
  - consistent enough to be used to create high-quality corpora to address specific linguistic issues
  - Large enough to be used by NN(?)
- Use as a test corpus for Quality Assessment
- Delimitation issues in terms of MWE boundaries lower the annotator agreement -> Indicates the possibility of a potential MWE
- Inter-annotator agreement increased when annotators used the discussion feature of the platform while annotating

Further work

- Focus on the use of decision flowcharts while annotating
- Find out what is the right amount of necessary data to train or fine-tune NN systems on the MWE annotation task
- Use our high-quality level corpus to test the NN systems for Quality Assessment

Thank you for your attention!

Any Questions?



Emmanuelle Esperança-Rodier

Yacine Haddad

Fiorella Albasini

Emmanuelle.Esperanca-Rodier@univ-grenoble-alpes.fr
Fiorella.Albasini@etu.univ-grenoble-alpes.fr