EUROPHRAS 2015

29 June - 2 July 2015, Malaga, Spain

Workshop Proceedings

MULTI-WORD UNITS IN MACHINE TRANSLATION AND TRANSLATION TECHNOLOGIES MUMTTT2015

1-2 July 2015

Editors: Gloria Corpas Pastor, Johanna Monti, Violeta Seretan, Ruslan Mitkov

2016. Editions Tradulex, Geneva
© The editors.
ISBN: 978-2-9700736-9-7

ORGANISING COMMITTEE

Workshop Chairs

Gloria Corpas Pastor (University of Malaga, Spain)

Johanna Monti (Università degli Studi di Sassari, Italy)

Violeta Seretan (Université de Genève, Switzerland)

Ruslan Mitkov (University of Wolverhampton, United Kingdom)

Advisory Board

Dmitrij O. Dobrovol'skij (Russian Academy of Sciences, Russia)

Carlos Ramisch (Aix-Marseille University, France)

Michael Rosner (University of Malta, Malta)

Agata Savary (Université François Rabelais Tours, France)

Kathrin Steyer (Institut für Deutsche Sprache, Germany)

Organisers

Rosario Bautista Zambrana

Cristina Castillo Rodríguez

Hernani Costa

Isabel Durán Muñoz

Jorge Leiva Rojo

Gema Lobillo Mora

Pablo Pérez Pérez Míriam

Seghiri Domínguez

Cristina Toledo Báez

Míriam Urbano Mendaña

Anna Zaretskaya

Secretary

Míriam Buendía Castro

Rut Gutiérrez Florido

Programme Committee

Iñaki Alegria (Universidad del País Vasco, Spain)

Giuseppe Attardi (Università di Pisa, Italy)

Doug Arnold (University of Essex, United Kingdom)

António Branco (Universidade de Lisboa, Portugal)

Paul Buitelaar (National University of Ireland, Galway)

František Čermák (Univerzita Karlova v Praze, Czech Republic)

Jean-Pierre Colson (Université Catholique de Louvain, Belgium)

Matthieu Constant (Université Paris-Est, France)

Gaël Dias (Université de Caen Basse-Normandie, France)

Mike Dillinger (Association for Machine Translation in the Americas)

Dmitrij O. Dobrovol'skij (Russian Academy of Sciences, Russia)

Peter Ďurčo (Univerzita sv. Cyrila a Metoda v Trnave, Slovak Republic)

Marcello Federico (Fondazione Bruno Kessler, Italy)

Sabine Fiedler (Universität Leipzig, Germany)

Natalia Filatkina (Universität Trier, Germany)

Thierry Fontenelle (Translation Centre for the Bodies of the European Union, Luxembourg)

Corina Forăscu (Al.I. Cuza University of Iasi, Romania)

Thomas François (Université Catholique de Louvain, Belgium)

Ulrich Heid (Universität Hildesheim, Germany)

Kyo Kageura (University of Tokyo, Japan)

Cvetana Krstev (University of Belgrade, Serbia)

Koenraad Kuiper (University of Canterbury, New Zealand)

Alon Lavie (Carnegie Mellon University, USA)

Malvina Nissim (Università di Bologna, Italy)

Preslav Nakov (Qatar Computing Research Institute, Qatar Foundation, Qatar)

Michael Oakes (University of Wolverhampton, United Kingdom)

Adriana Orlandi (Università degli studi di Modena e Reggio Emilia, Italy)

Yannick Parmentier (Université d'Orléans, France)

Pavel Pecina (Univerzita Karlova v Praze, Czech Republic)

Carlos Ramisch (Université de Grenoble, France)

Victoria Rosén (Universitetet i Bergen, Norway)

Michael Rosner (University of Malta)

Manfred Sailer (Goethe Universität, Germany)

Tanja Samardžić (Universität Zurich, Switzerland)

Agata Savary (Université François Rabelais Tours, France)

Gerold Schneider (Universität Zurich, Switzerland)

Gilles Sérasset (Université de Grenoble, France)

Max Silberztein (Universié de Franche-Comté, France)

Kiril Simov (Bulgarian Academy of Sciences, Bulgaria)

Kathrin Steyer (Institut für Deutsche Sprache, Germany)

Joanna Szerszunowicz (University of Bialystok, Poland)

Marko Tadić (University of Zagreb, Croatia)

Amalia Todirascu (Université de Strasbourg, France)

Beata Trawinski (Institut für Deutsche Sprache Mannheim, Germany)

Dan Tufiş (Romanian Academy, Romania)

Agnès Tutin (Université de Grenoble, France)

Lonneke van der Plas (Universität Stuttgart, Germany)

Veronika Vincze (University of Szeged, Hungary)

Martin Volk (Universität Zurich, Switzerland)

Eric Wehrli (Université de Genève, Switzerland)

Michael Zock (Aix-Marseille Université, France)

INVITED SPEAKER

Kathrin Steyer

Dr. Kathrin Steyer is Research Assistant at the Institute of German Language in Mannheim, the central state-aided institution for the study and documentation of the contemporary usage and recent history of the German language. K. Steyer received her PhD in German Studies/ General Linguistics at the University of Mannheim. She is head of the IDS Project *Usuelle Wortverbindungen* (multi-word expressions). From 2008-2010, she managed the German part of the multilingual EU project *SprichWort*. *Eine Internetplattform für das Sprachenlernen* (Proverb. An Internet Platform for Foreign Language Acquisition). In 2014, she was elected as president of EUROPHRAS. She is Co-editor of the *Yearbook of Phraseology* (de Gruyter/ Mouton) and a member of the review board of *Linguistik online*. Kathrin Steyer is author of numerous publications in the fields of theoretical phraseology and paremiology, corpus linguistics, multi-word pattern research, also in contrast with other languages, electronic lexicography; for example the monograph on a corpus-based theory of multi word pattern (cf. Steyer 2013) and the OWID Dictionary of German Proverbs; furthermore, Kathrin Steyer was the editor of the IDS Yearbook on phrases more or less fixed (2004) (de Gruyter) and the anthology of modern paremiology (2012), (both Narr Tübingen). Kathrin Steyer's project is a place to go for colleagues from all over the world who are interested in corpus-based studies phraseology, phraseography and lexicology, and the qualitative adaption of corpus analysis methods, especially of the collocation and pattern analysis.

MUMTTT2015 Academic Programme

Wednesday, 1 July 2015: 15.30-17:00

Session 1

Automatic extraction of multilingual MWU resources

Chaired by Johanna Monti

Bilingual Term Alignment for Language Service in Comparable Corpus
Zhiyan Zhao; Xuri Tang

Building a Lexical Bundle Resource for MT
Natalia Grabar; Marie-Aude Lefer

Assessing WordNet for Bilingual Compound Dictionary Extraction
Carla Parra Escartín; Héctor Martínez Alonso

Wednesday, 1 July 2015: 17:20-18:50

Session 2

Identification, acquisition and evaluation of multi-word terms

Chaired by Violeta Seretan

Multiword Units Translation Evaluation in Machine Translation: another pain in the neck?
Johanna Monti; Amalia Todirascu

Identifying Multi-Word Expressions from Parallel Corpora with String-Kernels and Alignment Methods
Federico Sangati; Andreas van Cranenburgh; Mihael Arcan; Johanna Monti; Marco Turchi

Parallel Sentences Fragmentation
Sergey Potemkin; Galina Kedrova

Thursday, 2 July 2015: 9:00-10:00

Session 3

Multilingualism and MWU processing. MWUs and word alignment techniques

Chaired by Ismail El Maarouf

Statistical Measures to characterise MWUs involving 'mordre' in French or 'bite' in English
Ismail El Maarouf; Michael Oakes

Aligning Verb+Noun Collocations to Improve a French - Romanian FSMT System
Amalia Todirascu; Mirabela Navlea

Thursday, 2 July 2015: 12:20-13:30

Session 4

Learning semantic information about MWUs from monolingual, parallel or comparable corpora

Chaired by Eric Wehrli

Aranea: Comparable Gigaword Web Corpora
Vladimír Benko; Peter Ďurčo

In-depth Study of the Phraseological Units in Islamic and Christian Religions in Samples (corpora) of Religious Texts
Madian Souliman; Ali Ahmad

Thursday, 2 July 2015: 15:00-16:30

Session 5

MWUs in Machine Translation

Chaired by Amalia Todirascu

Multiword Expressions in Machine Translation: The case of German compounds
Maria Ivanova; Eric Wehrli; Luka Nerima

Adding Multi-Word Expressions More Than Once in Machine Translation Liling Tan; Josef van Genabith

Multi-word Expressions in user-generated Content: How many and how well translated? Evidence from a Post-editing Experiment
Violeta Seretan

Thursday, 2 July 2015: 17:00-18:30

Session 6

Lexical, syntactic, semantic and translational aspects in MWU representation

Chaired by Carla Parra Escartín

Transformation and MWU in Quechua
Maximiliano Durán

Populating a Lexicon with Multiword Expressions in view of Machine Translation
Voula Gioli

MWEs: Light verb constructions vs Frozen constructions in Modern Greek and French
Angeliki Fotopoulou, Voula Giouli

Poster session

Challenges on the Translation of Collocations
Angela Costa

Chunking-based detection of English nominal compounds
Gábor Csernyi

Integrating Multi-Word Expressions in Statistical Machine Translation
Zied Elloumi, Laurent Besacier, Olivier Kraif

Analysis of Multiword Expression translation errors in Statistical Machine Translation
Natalia Klyueva, Jeevanthi Liyanapathira

Table of Contents

Keynote talk

Session 1 - Automatic extraction of multilingual MWU resources

Session 2 - Identification, acquisition and evaluation of multi-word terms

Session 3 - Multilingualism and MWU processing. MWUs and word alignment techniques

Session 5 - MWUs in Machine Translation

Session 6 - Lexical, syntactic, semantic and translational aspects in MWU representation

Poster session

Preface by the Workshop Chairs

This volume documents the proceedings of the 2nd Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2015), held on 1-2 July 2015 as part of the EUROPHRAS 2015 conference: "Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives" (Málaga, 29 June – 1 July 2015). The workshop was sponsored by European COST Action PARSing and Multi-word Expressions (PARSEME) under the auspices of the European Society of Phraseology (EUROPHRAS), the Special Interest Group on the Lexicon of the Association for Computational Linguistics (SIGLEX), and SIGLEX's Multiword Expressions Section (SIGLEX-MWE). The workshop was co-chaired by Gloria Corpas Pastor (Universidad de Málaga), Ruslan Mitkov (University of Wolverhampton), Johanna Monti (Università degli Studi di Sassari), and Violeta Seretan (Université de Genève). It received the support of the Advisory Board, composed of Dmitrij O. Dobrovol'skij (Russian Academy of Sciences, Moscow), Kathrin Steyer (Institut für Deutsche Sprache, Mannheim), Agata Savary (Université François Rabelais Tours), Michael Rosner (University of Malta), and Carlos Ramisch (Aix-Marseille Université).

The topic of the workshop was the integration of multi-word units in machine translation and translation technology tools. In spite of the recent progress achieved in machine translation and translation technology, the identification, interpretation and translation of multi-word units still represent open challenges, both from a theoretical and from a practical point of view. The idiosyncratic morpho-syntactic, semantic and translational properties of multi-word units poses many obstacles even to human translators, mainly because of intrinsic ambiguities, structural and lexical asymmetries between languages, and, finally, cultural differences. After a successful first edition held in Nice on 3 September 2013 as part of the Machine Translation Summit XIV, the present edition provided a forum for researchers working in the fields of Linguistics, Computational Linguistics, Translation Studies and Computational Phraseology to discuss recent advances in the area of multi-word unit processing and to coordinate research efforts across disciplines.

The workshop was attended by 53 representatives of academic and industrial organisations. The programme included 11 oral and 4 poster presentations, and featured an invited talk by Kathrin Steyer, President of EUROPHRAS. We received 23 submissions, hence the MUMTTT 2015 acceptance rate was 65.2%. The papers accepted are indicative of the current efforts of researchers and developers who are actively engaged in improving the state of the art of multi-word unit translation.

We would like to thank all authors who contributed papers to this workshop edition and the Programme Committee members who provided valuable feedback during the review process. We would also like to acknowledge the support and help from PARSEME and wish to thank in particular Agatha Savary for her collaboration. Finally, we would like to thank the local organisers, in particular Miriam Buendía and Rut Gutiérrez Florido, for all their work and their effort in the organisation of the workshop.

Gloria Corpas Pastor, Universidad de Málaga
Johanna Monti, University of Naples "L'Orientale"
Violeta Seretan, Université de Genève
Ruslan Mitkov, University of Wolverhampton

Corpus-driven Description of Multi-word Patterns

Kathrin Steyer
Institut für Deutsche Sprache
Mannheim
steyer@ids-mannheim.de

Abstract

This paper presents our model of 'Multi-Word Patterns' (MWPs). MWPs are defined as recurrent frozen schemes with fixed lexical components and productive slots that have a holistic – but not necessarily idiomatic – meaning and/or function, sometimes only on an abstract level. These patterns can only be reconstructed with corpus-driven, iterative (qualitative-quantitative) methods. This methodology includes complex phrase searches, collocation analysis that not only detects significant word pairs, but also significant syntagmatic cotext patterns and slot analysis with our UWV Tool. This tool allows us to bundle KWICs in order to detect the nature of lexical fillers for and to visualize MWP hierarchies.

First we discuss the nature of MWPs as frozen communicative units. Then, we illustrate our methodology and selected linguistic results using examples from a contrastive study of German, Spanish, and English prepositional MWPs.

1 Introduction

Learning from corpora does not just mean to find a certain number of similar citations that confirm a hypothesis. It means knowledge about patterns of language use. Patterns can be reconstructed from corpus analysis by collecting many similar use cases – bottom up in a corpus-driven way. Looking at many use cases does not mean describing what is already known and visible: It means seeing hidden structures. This is not merely 'more data', but new interrelations, unusual cross-connections, surprising relationships, and networks. Of course, pattern detection is not a new invention but one of the central methods in information science, data mining, and information retrieval. But, we are convinced that in respect to a qualitative reconstruction of hidden patterns in language use and their applications in lexicography and second language teaching, we are just at the beginning. We would like to discuss this pattern view of language use on the basis of multi-word expressions and phrasemes.

2 Multi-word Patterns

Due to the rise of corpus linguistics and the feasibility of studying language data in new quantitative dimensions, it became more and more evident that language use is fundamentally made up by fixed lexical chunks, set phrases, long-distance word groups, and multi-word expressions (MWEs). Sinclair's inductively reconstructed collocations (cf. 1991) and Hausmann's collocation pairs (cf. 2004) are the two leading concepts in collocation research. Basically, they are merely different ways of looking at the same fundamental principle of language: linguistic frozenness and fixedness. Compositional collocations and idioms differ in their degree of lexical fixedness and semantic opacity, their recognisability and prototypicality (cf. Moon 1998, Burger et al. 2007). But they all share the most important characteristic: They are congealed into autonomous units in order to fill a specific role in communication. All these fragments are fixed patterns of language use (cf. Hunston/Francis 2000; cf. Hanks 2013). There is no core and no periphery. The difference is only in the degree of conspicuousness for the observer. These word clusters did not become fixed expressions by chance, but because there was a need of speakers for an economic way of communicating (cf. Steyer 2013). Currently, this widening of scope to every kind of frozen multi-word unit is also accepted in modern phraseology, as Dobrovol'skij outlined in 2011 in the third volume of "Konstruktionsgrammatik" in a very compact way.

Lately, not only multi-word research but also usage-based linguistics as a whole is subject to a shift. If you conduct empirical studies on corpora systematically and – this is very important – in a bottom up way, it is evident that MWEs are not as singular and unique as it is often still assumed in phraseology. MWEs are linked in many ways with other units in the lexicon. They are specific lexical realisations of templates, definitely more noticeable and more fixed than ad-hoc formulations, but not unique. Such templates emerge from repeated usage and can be filled with ever changing lexical elements, both phraseological and non-phraseological. We call them 'Multi-word Patterns' (MWPs) (cf. Steyer 2013)[1].

MWPs are recurrent frozen schemes with fixed lexical components and productive slots that have holistic – but not necessarily idiomatic – meanings or functions, sometimes only on an abstract level. The slots are filled with lexical units that have similar lexical-semantic and/or pragmatic characteristics, but must not belong to the same morpho-syntactic class. Speakers are able to recall those schemes as lexicon entries and fill the gaps in a specific communicative situation in a functionally adequate way. For example, the sentence *Die Worte* ***klingen fremd für westliche Ohren*** (*The words* ***sound strange for Western ears***) is based on the following MWP:

(1)
für **X** ***Ohren*** Y ***klingen***
(ww: to sound Y for X ears)

X ADJ$_{\{\text{HUMAN}\}}$ fillers: *deutsche* (German) / *westliche* (Western) */ europäische* (*European*) /…

Y ADV{$_{\text{CONNOTATION}}$} fillers: *fremd* (foreign) / *ungewohnt* (unfamiliar) / *exotisch* (exotic) / *seltsam* (strange) / *vertraut* (familiar) / *merkwürdig* (odd) / *schräg* (discordant) / *pathetisch* (melodramatic) /...

Holistic Meaning:
'Somebody (a person / a group of people / a specific community) could possibly perceive, interpret, or assess something in a certain way'

The X ADJ fillers refer to a person, to groups of people, or to specific communities. The Y ADV collocations are almost always connotative adverbs. The whole pattern expresses specific interpretations of a fact or situation. But the speaker do not present the interpretation or evaluation as his own. He pretends that this is the interpretation of an abstract or fictional group of people. So the speaker can present the interpretation as possible or given without having to take responsibility for it.

MWEs and MW patterns are not clear-cut or distinct entities. On the contrary, fragments and overlapping elements with fuzzy borders are typical for real language use. This means that there are rarely MWEs as such. In real communicative situations, some components are focused while others fade to the background.

The reconstruction of MWPs is only possible with complex corpus-driven methods in an iterative way (quantitative – qualitative).[2] Generally, we study the nature of MW patterns by exploring keyword-in-context concordances (KWIC) of multi-word units. Beside complex phrase searches and reciprocal analysis with COSMAS II (cf. CII), we use mainly two empirical methods for KWIC bundling: We assess collocation profiles that are calculated by the IDS collocation analysis algorithm (cf. Belica 1995). This type of collocation analysis bundles KWICs and citations according to the LLR (log likelihood ratio) and also summarizes the results as lists of collocation clusters and syntagmatic patterns (compare Figure 1 in 3.). The second method is exploring and bundling KWICs with our UWV Tool that allows us to define search patterns with specific surface characteristics, depending on our research question or hypothesis (cf. Steyer/Brunner 2014). The search patterns are essentially regular expressions consisting of fixed lexical items and gaps between those (with an arbitrary length, i.e., the fillers do not have to be single words, but can also be n-grams). The fillers are ranked according to frequency, and it is also possible to annotate them with tags, to add narrative comment, and to output annotated filler groups. All this interpreted data can be exported for a lexicographic online representation, recently as "Multi-Word Fields" (cf. Steyer et al. 2013).

[1] This term is similar to the term 'phrasem-constructions' proposed by Dobrovols'kij in 2011. But we prefer Steyer's term because we do not want to focus on the construction grammar framework, but take a strictly lexical and first and foremost usage-based perspective. Without doubt, the discussion of the relationship between these approaches is high on our agenda.

[2] The following examples are all taken from the *German Reference Corpus* (*Deutsches Referenzkorpus*) (cf. DeReKo), located at the Institut for the German Language in Mannheim. Our focus lies on syntagmatic word surface structures, and we use corpora that are not morpho-syntactically annotated.

In the following chapter, we illustrate our methodology and selected linguistic results using examples from a new contrastive project (German – Slovakian – Spanish)[3].

We concentrate on the German - Spanish contrast (with added English examples), but the main aspects can also be observed in Slovakian.

3 MWP in Contrast – Methods and Evidences

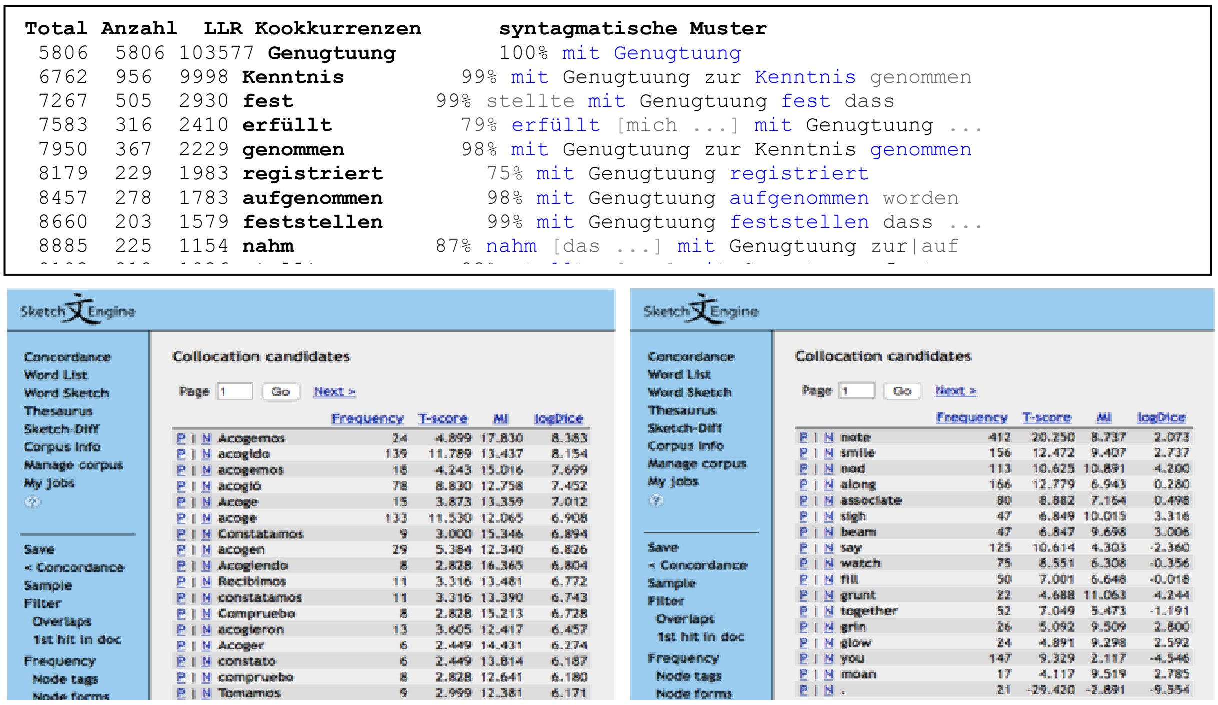
Our research goal is the detection and description of prepositional MWE and MW patterns like *nach Belieben* (at will), *mit Genugtuung* (with satisfaction), *am Ende* (at the end). We explore and describe their fixedness, variance, and usage on several levels of abstraction and in interlingual contrast.

The key questions are:

- On which level can we find differences in the use of prepositional MWE and patterns in the three languages?
- Are there parallels on higher levels of abstraction that allow us to assume universal functional concepts?
- Is it possible to visualize these relationships and if yes, which kind of representation is appropriate for which audience, for example for foreign language acquisition?

The following two aspects are in the center of our multilingual analysis: a) collocation fields in contrast and b) lexical filler and cotext patterns in contrast. We will now look at the MWP *mit Genugtuung* (*con satisfacción / with satisfaction)* as an example.

With the help of collocation profiles calculated with CA for German and with Sketch Engine for Spanish and English (see Figure 1) we describe the meaning and usage and identify phenomena of convergence and divergence:



```
Total Anzahl  LLR Kookkurrenzen       syntagmatische Muster
 5806  5806 103577 Genugtuung          100% mit Genugtuung
 6762  956  9998 Kenntnis            99% mit Genugtuung zur Kenntnis genommen
 7267  505  2930 fest              99% stellte mit Genugtuung fest dass
 7583  316  2410 erfüllt             79% erfüllt [mich ...] mit Genugtuung ...
 7950  367  2229 genommen            98% mit Genugtuung zur Kenntnis genommen
 8179  229  1983 registriert           75% mit Genugtuung registriert
 8457  278  1783 aufgenommen           98% mit Genugtuung aufgenommen worden
 8660  203  1579 feststellen           99% mit Genugtuung feststellen dass ...
 8885  225  1154 nahm              87% nahm [das ...] mit Genugtuung zur|auf
```

	Frequency	T-score	MI	logDice
P | N Acogemos	24	4.899	17.830	8.383
P | N acogido	139	11.789	13.437	8.154
P | N acogemos	18	4.243	15.016	7.699
P | N acogió	78	8.830	12.758	7.452
P | N Acoge	15	3.873	13.359	7.012
P | N acoge	133	11.530	12.065	6.908
P | N Constatamos	9	3.000	15.346	6.894
P | N acogen	29	5.384	12.340	6.826
P | N Acogiendo	8	2.828	16.365	6.804
P | N Recibimos	11	3.316	13.481	6.772
P | N constatamos	11	3.316	13.390	6.743
P | N Compruebo	8	2.828	15.213	6.728
P | N acogieron	13	3.605	12.417	6.457
P | N Acoger	6	2.449	14.431	6.274
P | N constato	6	2.449	13.814	6.187
P | N compruebo	8	2.828	12.641	6.180
P | N Tomamos	9	2.999	12.381	6.171

	Frequency	T-score	MI	logDice
P | N note	412	20.250	8.737	2.073
P | N smile	156	12.472	9.407	2.737
P | N nod	113	10.625	10.891	4.200
P | N along	166	12.779	6.943	0.280
P | N associate	80	8.882	7.164	0.498
P | N sigh	47	6.849	10.015	3.316
P | N beam	47	6.847	9.698	3.006
P | N say	125	10.614	4.303	-2.360
P | N watch	75	8.551	6.308	-0.356
P | N fill	50	7.001	6.648	-0.018
P | N grunt	22	4.688	11.063	4.244
P | N together	52	7.049	5.473	-1.191
P | N grin	26	5.092	9.509	2.800
P | N glow	24	4.891	9.298	2.592
P | N you	147	9.329	2.117	-4.546
P | N moan	17	4.117	9.519	2.785
P | N .	21	-29.420	-2.891	-9.554

Figure 1. Collocation profiles *mit Genugtuung* (cf. CA) and *con satisfacción – with satisfaction* (cf. SKE) (snippet)

[3] Three partner institutions are involved in this research project: the University of Santiago de Compostela (Head: Carmen Mellado Blanco), the University of Trnava (Head: Peter Ďurčo), and the IDS with the UWV research group (Head: Kathrin Steyer).

The collocation profiles give us strong evidence for a restriction of verbal collocation partners: This multi-word expression is prototypically combined with verbs that refer to communicative acts:

(2)
[*mit Genugtuung* V]:
  V partners: *mitteilen / sagen / hinweisen / ankündigen / zur Kenntnis nehmen /…*
[*con satisfacción* V]
  V partners: *constatar* (to be stated) / *recibir* (to admire) / *saludar* (to appreciate) / *observar* (to observe) /…
[*with satisfaction* V]:
  V partners: *note / say / remark / reflect /…*

Because of the verbal convergence, you can assume an interlingual abstract pattern:

[*mit / con / with* SUB{$_{\text{EMOTION}}$} V{$_{\text{COMMUNICATION}}$}]

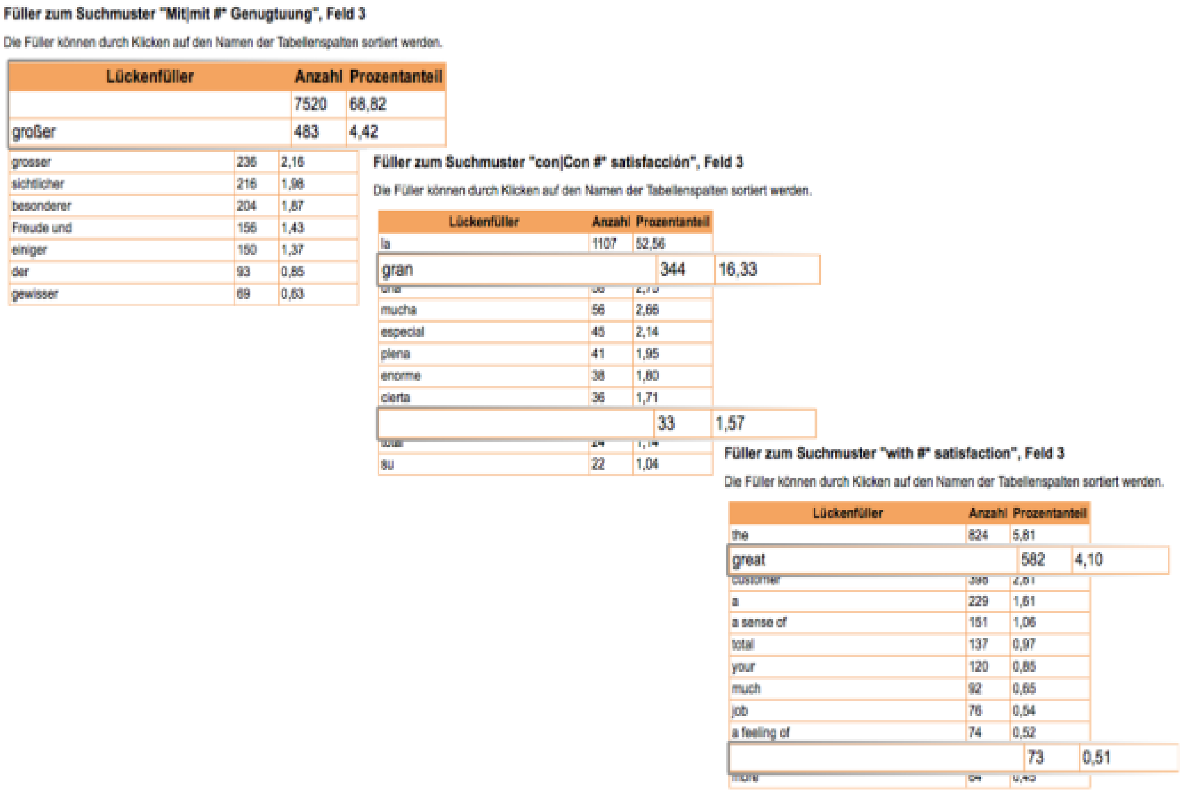
An interesting difference can be observed between German, on the one hand, and Spanish and English on the other hand: Many verbal collocation partners on the highest ranks of the Spanish *con satisfacción* and the English *with satisfaction* refer to non-verbal behavior like *nod / smile / beam / grunt* resp. *reír* (to laugh) / *sonreír* (to smile) / *suspirar* (to sigh) / *respirar* (to breathe) / *fruncir los labios* (to purse one's lips). In German, this kind of contextualization is a very rare phenomenon.

In a second step, we generate filler tables with the help of our UWV Tool and compare them between the languages (see Figure 2):

First of all, the tables give information concerning the degree of lexical fixedness. As Figure 2 shows, the gap between the preposition *mit* and the noun *Genugtuung* is empty in approx. 70% of occurrences. This "Zero Gap" indicates a high degree of lexicalization and a lexicon entry *mit Genugtuung*. In Spanish and English, this empty slot is not so recurrent. Instead a strong internal variance is established.

In all three tables, we can observe groups of ADJ fillers with the same communicative functions: a) intensification, e.g., *groß – gran – great*, and b) connotation, e.g., *grimmiger – insana* (insane) – *grim.* In many cases, both functions overlap.

As mentioned in Chapter 2, the UWV tool enables to define any size of slots. Figure 3 (see next page) illustrates – for example – typical trigram fillers in German, Spanish and English.



Füller zum Suchmuster "Mit|mit #" Genugtuung", Feld 3

Die Füller können durch Klicken auf den Namen der Tabellenspalten sortiert werden.

Lückenfüller	Anzahl	Prozentanteil
	7520	68,82
großer	483	4,42
grosser	236	2,16
sichtlicher	216	1,98
besonderer	204	1,87
Freude und	156	1,43
einiger	150	1,37
der	93	0,85
gewisser	69	0,63

Füller zum Suchmuster "con|Con #" satisfacción", Feld 3

Die Füller können durch Klicken auf den Namen der Tabellenspalten sortiert werden.

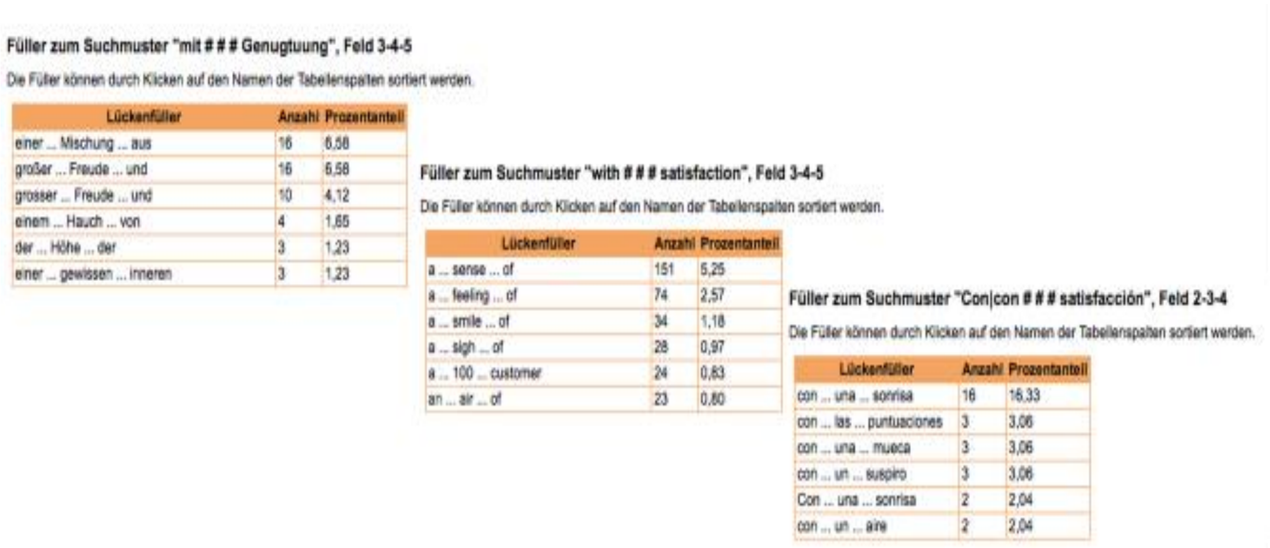
Lückenfüller	Anzahl	Prozentanteil
la	1107	52,56
gran	344	16,33
una	[illegible]	[illegible]
mucha	56	2,66
especial	45	2,14
plena	41	1,95
enorme	38	1,80
cierta	36	1,71
	33	1,57
total	24	[illegible]
su	22	1,04

Füller zum Suchmuster "with #" satisfaction", Feld 3

Die Füller können durch Klicken auf den Namen der Tabellenspalten sortiert werden.

Lückenfüller	Anzahl	Prozentanteil
the	824	5,81
great	582	4,10
customer	[illegible]	[illegible]
a	229	1,61
a sense of	151	1,06
total	137	0,97
your	120	0,85
much	92	0,65
job	76	0,54
a feeling of	74	0,52
	73	0,51
more	64	[illegible]

Figure 2. Filler tables of *mit – con –* with # (1 slot) *Genugtuung – satisfacción – satisfaction* (snippet)



Füller zum Suchmuster "mit # # # Genugtuung", Feld 3-4-5

Die Füller können durch Klicken auf den Namen der Tabellenspalten sortiert werden.

Lückenfüller	Anzahl	Prozentanteil
einer ... Mischung ... aus	16	6,58
großer ... Freude ... und	16	6,58
grosser ... Freude ... und	10	4,12
einem ... Hauch ... von	4	1,65
der ... Höhe ... der	3	1,23
einer ... gewissen ... inneren	3	1,23

Füller zum Suchmuster "with # # # satisfaction", Feld 3-4-5

Die Füller können durch Klicken auf den Namen der Tabellenspalten sortiert werden.

Lückenfüller	Anzahl	Prozentanteil
a ... sense ... of	151	5,25
a ... feeling ... of	74	2,57
a ... smile ... of	34	1,18
a ... sigh ... of	28	0,97
a ... 100 ... customer	24	0,83
an ... air ... of	23	0,80

Füller zum Suchmuster "Con|con # # # satisfacción", Feld 2-3-4

Die Füller können durch Klicken auf den Namen der Tabellenspalten sortiert werden.

Lückenfüller	Anzahl	Prozentanteil
con ... una ... sonrisa	16	16,33
con ... las ... puntuaciones	3	3,06
con ... una ... mueca	3	3,06
con ... un ... suspiro	3	3,06
Con ... una ... sonrisa	2	2,04
con ... un ... aire	2	2,04

Figure 3. Trigram filler tables *mit – with – con # # # Genugtuung – satisfaction – satisfacción* (snippet)

Interesting phenomena of convergence are recurrent evaluative quantifier or intensifier phrases in all three languages:

(3)
[*mit* X *Genugtuung*]
X fillers: *einem Hauch von / einem Anflug von / einem Schuss von / einer Prise von /...*
[*con* X *satisfacción*]
X fillers: *mayor nivel de* (higher level of) */ un grado de* (a degree of) */ una pizca de* (a pinch of) / *algún grado de* a degree (some degree of) */...*
[*with* X *satisfaction*]
X fillers: *a sense of / a feeling of / a great deal of /...*

This suggests an interlingual tendency to express a scale of satisfaction in a more or less indirect way.

Another example of convergent bigram fillers are coordinative structures, e.g., appositions of nouns with positive connotations.

(4)
[*mit* X ***und*** *Genugtuung*]
N fillers: *Stolz / Freude* / *Häme* (scorn)
[*con* X ***y*** *satisfacción*]
N fillers: *orgullo* / *alegría* /*asombro* (wonder)
[*with* X *and satisfaction*]
N fillers: *pride* / *joy* / *pleasure*

Because of these pronounced similarities we assume universal abstract patterns with a holistic function of intensification.

(5)
[*mit* N$_{\{\mathrm{EMOTION}\}}$ + *und* + *Genugtuung*]
[*con* N$_{\{\mathrm{EMOTION}\}}$ + *y* + *satisfacción*]
[*with* N$_{\{\mathrm{EMOTION}\}}$ + *and* + *satisfaction*]

[*mit* N$_{\{\mathrm{EMOTION}\}}$ + *und* + N$_{\{\mathrm{EMOTION}\}}$]
[*con* N$_{\{\mathrm{EMOTION}\}}$ + *y* + N$_{\{\mathrm{EMOTION}\}}$]
[*with* N$_{\{\mathrm{EMOTION}\}}$ + *and* + N$_{\{\mathrm{EMOTION}\}}$]

[P + N$_{\{\mathrm{EMOTION}\}}$ + *und / y / and* + N$_{\{\mathrm{EMOTION}\}}$]

In our lexicographic description, we will try to show convergences and divergences between the languages with the aid of collocation fields and slot-filler tables on several levels of abstraction. These will be annotated and systematized according to typical usage characteristics and linked across languages.

4 Conclusion

If patterns and imitation are the genuine principles of language production and reception, they must move to the focus of lexicographic description, language acquisition, and machine translation. How these highly complex, overlapping phenomena can be structured and explained in a didactically effective way will be one of the most exciting questions for future researches in these fields.

Reference

Belica, Cyril 1995. *Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode.* Institut für Deutsche Sprache, Mannheim.

Burger, Harald, Dmitrij Dobrovol'skij, Peter Kühn and Neal R. Norrick. 2007 (eds.). *Phraseologie. Ein internationales Handbuch zeitgenössischer Forschung/Phraseology. An international Handbook of Contemporary Research.* (2 Editions). (HSK 28, 1/2). de Gruyter, Berlin/New York:.

Dobrovol'skij, Dmitrij. 2011. Phraseologie und Konstruktionsgrammatik. In Lasch, Alexander and Alexander Ziem (eds.), *Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze*. (Stauffenburg Linguistik 58). Stauffenburg, Tübingen: 111-130.

Hanks, Patrick. 2013. *Lexical Analysis. Norms and Exploitations*. The MIT Press, Cambridge, MA /London.

Hausmann, Franz Josef 2004. 'Was sind eigentlich Kollokationen?' In Steyer, Kathrin (ed.), *Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003*. de Gruyter, Berlin/New York: 309-334.

Hunston, Susan and Gill Francis. 2000. *Pattern Grammar. A corpus-driven approach to the lexical grammar of English.* Amsterdam/Philadelphia: John Benjamins.

Moon, Rosamund 1998. *Fixed Expressions and idioms in English. A Corpus-Based Approach*. Clarendon Press, Oxford.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Steyer, Kathrin. 2013. *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. (Studien zur Deutschen Sprache 60). Narr, Tübingen.

Steyer, Kathrin and Annelen Brunner. 2014: Contexts, Patterns, Interrelations - New Ways of Presenting Multi-word Expressions.
Proceedings of the 10th Workshop on Multiword Expressions (MWE). ACL Anthology. http://www.aclweb.org/anthology/W/W14/W14-0814.pdfshttp://www.aclweb.org/anthology/W/W14/W14-0814.pdf

Internet Sources (Accessed on 7 August 2015)

CII 2015. *COSMAS II. Corpus Search, Management and Analysis System* http://www.ids-mannheim.de/cosmas2/

DeReKo 2015. *Deutsches Referenzkorpus. / Archiv der Korpora geschriebener Gegenwartssprache 2014-II* (Release 11.09.2014). Institut für Deutsche Sprache, Mannheim: http://www.ids-mannheim.de/DeReKo.

SKE: Sketch Engine. https://www.sketchengine.co.uk/

Steyer, Kathrin, Annelen Brunner and Christian Zimmermann. 2013. Wortverbindungsfelder Version 3: Grund. http://wvonline.ids-mannheim.de/wvfelder-v3/

Appendix:

Abbreviations

ADJ:	adjective
ADV:	adverb
CA:	IDS collocation analysis (Belica 1995)
MWE:	multi word expression
MWP:	multi word pattern
N:	noun
UWV:	Usuelle Wortverbindungen
V:	verb

Figures

Assessing WordNet for Bilingual Compound Dictionary Extraction

Carla Parra Escartín
Hermes Traducciones
Madrid, Spain
carla.parra@hermestrans.com

Héctor Martínez Alonso
University of Copenhagen
Copenhagen, Denmark
alonso@hum.ku.dk

Abstract

We present work to explore ways of automatically retrieving compound dictionaries from sentence-aligned corpora using WordNet. More specifically, we focus on the pair of languages German to Spanish and try to retrieve the Spanish translational correspondences of German nominal compounds. German compounds are a challenge because their correspondences into other languages are not straightforward and better methods for aligning them successfully to their translations in parallel corpora are needed. We describe a pilot experiment carried out to assess whether it is possible to align the parts of a German compound with the words in the Spanish translation which correspond to the main WordNet categories.

1 Introduction

As Sag et al. (2001) argue in their seminal "pain in the neck" paper, Multiword Expressions (MWEs) are a major bottleneck for many Natural Language Processing (NLP) applications. Within the MWE community, a lot of efforts have been made to automatically extract MWEs. As Ramisch (2012) points out, there has been a shift in the MWE research and now researchers focus also in integrating MWEs in applications. Our research had as a starting point a real problem for human translation and Machine Translation (MT), and therefore is application-driven. Both human translators and Machine Translation Systems encounter problems to translate German nominal compounds into other languages such as Spanish. Although we focus on compound dictionary extraction, the ultimate aim is to integrate the extracted dictionaries in Statistical Machine Translation (SMT) tasks.

The remainder of this paper is organised as follows. Section 2 discusses German compounds, why they are a challenge and why we treat them as MWEs. Section 3 offers an overview of strategies to retrieve compound dictionaries, and Section 4 presents the case-study we used in our pilot experiment. The results are discussed in Section 5. Section 6 summarises the paper and discusses future work.

2 German Compounds

One of the main issues when dealing with MWEs is that there does not exist a widely agreed upon typology (Moon, 1998; Cowie, 1998; Sag et al., 2001; Baldwin and Kim, 2010, etc.). With regard to German compounds, they could be considered MWEs written together in one typographic word. Example 1 below exemplifies this showing the inner structure of the German compound "*Straßenlampe*" and its translation into English.

(1) *Straße* *en* *Lampe*
street Ø lamp/light
[EN]: 'street lamp // street light'

Although "*Straßenlampe*" would usually be considered a nominal compound but not a MWE, its possible translational correspondences in English are nominal compounds which are considered MWEs. However, if we split the German compound as in Example 1, we can see that in fact it is formed by two nouns joined together by the filling element "*en*". Moreover, the translations of German nominal compounds into other languages usually correspond to phrases which again could be considered MWEs. This is illustrated in Example 2, where the different parts of the German compound "*Warmwasserbereitungsanlagen*" are shown together with its correspondences in English and Spanish.

(2)

Warm	*Wasser*	*Bereitung*	*s*	*Anlagen*
warm	water	production	Ø	systems
caliente	agua	producción	Ø	sistemas

[EN]: 'warm water production systems'

[ES]: 'sistemas de producción de agua caliente'

Based on the arguments given above, here we treat German compounds and their translations into Spanish as a special MWE problem.

3 Related work: Compound Dictionary Extraction

Compound dictionary extraction is not a usual NLP task. In fact, most researchers have rather focused on the automatic translation of compounds. Moreover, the majority of the approaches towards the automatic translation of compounds have focused on compounds of the type "noun-noun", and for other language pairs different than German and Spanish.

3.1 Dictionary-based approaches

Rackow et al. (1992) explored how to automatically retrieve the translation of nominal compounds for the pair of languages German-English. They restricted their study to compounds of the kind "noun-noun" arguing that this is the most common type of compounds in German. Upon splitting the compound, its parts were looked up in the lexicon of the MT system they were using. If the words did not appear in the lexicon, the online accessible Machine Readable Dictionary (MRD) Collins German-English was used. All the possible translations of the compound parts were looked up and stored. To filter the right translations, the candidates were checked against a monolingual corpus and the most frequent combinations were chosen. In the case of infrequent compounds, the compound modifiers were checked and the most frequent form was selected in each case.

In 1999, Grefenstette (1999) used a similar approach, but he used the web to filter out the translation candidates. He tested his approach with German nominal compounds and Spanish nominal phrases and in both cases tried to translate these into English. He reported to have retrieved 86% of Spanish to English translations, and 87% of German to English translations.

Moa (2005) tested the same approach for Norwegian-English using the search engines Google and Yahoo!, also with positive results.

Tanaka and Baldwin (2003b) also explored a similar method for translating Japanese noun-noun compounds into English. They used a word-level transfer dictionary, and a target language monolingual corpus. The candidate translations were ranked using what they called a "corpus-based translation quality" measure. This measure used the probabilities of the component parts and the probability of the candidate as a whole to select the translations.

Finally, Mathur and Paul (2009) also explored a similar method for the automatic translation of English nominal compounds into Hindi. They focused on noun-noun compounds in English. The novelty of their approach consisted of selecting the correct sense of the component parts by running a Word Sense Disambiguation (WSD) system on the source language data. Only the disambiguated word senses were looked up in the dictionary thus reducing the number of translation candidates to be filtered out. Translation candidates were selected by looking for a set of "translation templates" (patterns of possible word category combinations in Hindi when translating English noun-noun compounds).

More recently, Weller and Heid (2012) explored ways of aligning German compounds and English MWE terms using comparable corpora. First all compound parts were translated individually using a bilingual dictionary and then the translations were aligned to their English counterparts. Additionally, they analysed the English translations and derived a set of term/Part-of-Speech (PoS) patterns to be used in the alignment process. The usage of PoS patterns improved the overall precision of the alignment task.

3.2 Machine Learning approaches

Tanaka and Baldwin (2003a; 2003b) carried out a feasibility study on shallow methods to translate compound nouns in memory-based MT (MBMT) and word-to-word compositional dynamic MT (DMT) for Japanese and English noun-noun compounds. A year later, they used Support Vector Machines (SVMs) to rank the translation candidates (Baldwin and Tanaka, 2004). This machine learning method used a bilingual dictionary, syntactic templates for translation candidate generation, and corpus and dictionary statistics for selection.

Bungum and Oepen (2009) investigated the automatic translation of Norwegian nominal compounds. They proposed the usage of a conditional Maximum Entropy ranker to filter out the translation candidates.

3.3 Compound dictionary extraction: conclusion

The most common strategies for facing the translation of nominal compounds consist on using bilingual dictionaries. However, this approach relies on the availability of such dictionaries. In the case of German and Spanish, no freely available dictionary which could be used for NLP tasks was found. Our potential experiments were thus restricted by the resources available and we thought of ways of retrieving the translation of German nominal compounds into Spanish using only minimal linguistic information.

Here, we present a different approach using WordNet (Fellbaum, 1998). WordNet has been widely used as semantic inventory for Word Sense Disambiguation and many other NLP tasks. Several authors have investigated the semantic properties of nominal compounds using WordNet reporting positive results (Girju et al., 2005; Kim and Baldwin, 2013, etc). However, we did not identify experiments using WordNet to align the parts of a compound and its corresponding translations in another language.

4 Case Study

We take as a preliminary hypothesis that the different parts of a compound will share semantic features with their corresponding translational equivalents in other languages. Based on this hypothesis, we run a pilot experiment to verify whether it holds true.

Our pilot experiment consists on semantically tagging the parts of the compounds in German and their Spanish translations using WordNet and trying to find out possible overlappings across languages. We expected to be able to align the split German compound with the Spanish MWE by finding a correspondence between the semantic types of their parts. Example 3 below exemplifies this using as an example "*Handbremsvorrichtung*" (hand brake device).

(3) *Hand.BODY PART* *Bremse.ARTIFACT*
mano.BODY PART freno.ARTIFACT
Vorrichtung.ARTIFACT
dispositivo.ARTIFACT
[DE]: 'Handbremsvorrichtung'
[ES]: 'Dispositivo de freno de mano'

As can be observed in 3, the semantic types of the different compound parts happen to meet the semantic types of the content words of its translation into Spanish.

4.1 Gold Standard

We created a Gold Standard consisting of 168 German compounds and their translations. The data was extracted from two short files of the TRIS corpus (Parra Escartín, 2012), a specialised German-Spanish corpus. Only *1:n* cases (German compound → Spanish MWE) were taken into account. Compounds in coordination involving ellipsis an abbreviated compounds were disregarded.

All compounds were split using the compound splitter developed by Weller and Heid (2012) because according to the compound splitter comparison carried out in Parra Escartín (2014), this was the best compound splitter for German. For each compound in our Gold Standard, we had its parts, lemmatised and tagged with their corresponding PoS. On the Spanish side, all translation correspondences were also lemmatised and PoS tagged. The Spanish corpus was tagged using the TreeTagger PoS tagger (Schmid, 1994). Whenever a compound had several translation correspondences, each of them was stored as a different entry in the Gold Standard.

4.2 Experiments

As we were only aiming at carrying out a feasibility study, for our experiments we used as our corpus the two short texts we had used to extract the Gold Standard. Similarly to what we did to the Gold Standard, we lemmatised and PoS tagged the corpus. Then, we tagged it semantically using both WordNets. We assumed that the senses in each WordNet were ordered hierarchically, and thus took the first sense appearing for each word as its semantic type. No WSD was performed.

A prior semantic matching between the German and the Spanish WordNet was needed. This semantic type matching had to be done manually, and the main challenge faced consisted on a mismatch between the data types in German and Spanish. There are *n:n* and *n:1* correspondences because GermaNet (Hamp and Feldweg, 1997) and the Spanish Wordnet (Gonzalez-Agirre et al., 2012) do not share their semantic types. GermaNet has different semantic types for adjectives, whereas the Spanish WordNet does not. As a result, we had an uneven semantic type matching accross both WordNets.

Once there had been established a way of aligning the data types across the two WordNets, we carried out a compound-phrase matching task. This task was carried out automatically given the following premises:

1. Given a split German compound C, there is a list of lemmas $C = [c_0, ..., c_n]$.
2. Given a Spanish sentence aligned to the German sentence that contains C, there is a list of lemmas $S = [s_0, ..., s_n]$.
3. Be *type*(x) a function that retrieves the semantic type of a word, obtained from WordNet.
4. For each German compound, Spanish sentence pair (C,S):
    1. Locate the translated root of C in S by finding a lemma s_x in S with a semantic type that matches the root of the compound, i.e. $type(s_x) = type(c_n)$.
    2. From this point, two strategies were tested:
       Span: Locate the rightmost word in the Spanish phrase that translates C by finding a lemma s_y in S with a semantic type that matches the first lemma of the compound, i.e. $type(s_y) = type(c_0)$. The candidate Spanish phrase that translates C is the span of words defined as $[s_x, ..., s_y]$.
       Size: Starting from s_x, s_y is the seventh word starting from s_x. When s_y is a stopword, the size of the sequence is reduced up to the previous lexical word in the sentence and the value s_y is then updated to that word. The candidate Spanish phrase that translates C is the span of words defined as $[s_x, ..., s_y]$.

5 Results

Once the experiment had finished, we tested whether our hypothesis held for our Gold Standard. Unfortunately, only one candidate translation was retrieved, and it was wrong. This candidate was found in the so-called "Size" approach in Section 4.2.

As the experiments did not retrieve any results we could analyse, we analysed our Gold Standard to determine possible sources of error. First, we checked that all the entries in our Gold Standard could be semantically tagged using WordNet. The results of our evaluation are shown in Table 1 below.

	No. of items	[%]
Total Pairs	133	100%
Perfect coverage pairs	74	55.6%
Perfect coverage German	39	29.3%
Perfect coverage Spanish	9	6.8%
WN coverage error on both	11	8.3%

Table 1: Results obtained after running our pilot experiment.

As may be observed in Table 1, one of the main issues we faced was the WordNet coverage. If a word could not be found in WordNet, the subsequent steps in the experiment failed. This may be partially due to noise introduced by the PoS taggers. In fact, the Spanish PoS tagging was damaging the experiment, particularly when the heads of the compound could not be retrieved. We also encountered cases in which a noun was tagged as a verb. For instance, the noun *importes* (EN: amounts), was tagged as a form of the verb *importar* (EN: to matter). When these errors occur, the looking up process in WordNet fails, as we are looking up a completely different word. In German, there were also a couple of lemmatisation errors that prevented that words present in WordNet were successfully looked up. The words *Anlagen* and *Familien* (EN: systems and families, respectively) were in some compounds not lemmatised, whereas in other compounds they were correctly lemmatised to

Anlage and *Familie*. This led to further errors in the semantic tagging.

Second, we realised that we may have also introduced sources of error in our semantic type mapping. As discussed, earlier, GermaNet has useful additional information for German which maps unevenly to other WordNets. In Spanish, we deleted the differences in semantic types across categories to allow for a more general mapping between the different languages. Before running further experiments, our mapping shall be revised, particularly in the case of verbs.

Third, we realised that our hypothesis was not holding for the 74 pairs that we had managed to semantically tag on both languages. In fact, only 13 pairs were having the exact same match as we had expected. In one case (*Darlehensförderung*, EN: loan promotion), the tags were right, but the translation was not reversing the order of the elements as we had expected. However, upon revising this translation into Spanish, it seems that it was a translation mistake done by the translator of the text. In another case, the match does not work because we have a *2:3* alignment. *Umweltenergie* (EN: environmental energy), is translated into Spanish as *energía del medio ambiente*. This is because *Umwelt* (EN: environment) can be translated into Spanish as the nominal compound *medio ambiente*. Such cases had not been foreseen either.

Fourth, we also realised that in some cases a WSD task would have increased the number of matches. An alternative would have been to look up all senses for each word and try to find matches across the list of senses for each word in each language. However, as our semantic type matching was *1:1*, this would have been problematic.

Finally, we also realised that in GermaNet the senses seem not to be hierarchically ordered, which poses an additional problem, as we only used the first sense in each WordNet for our experiment. For instance, the first sense of the German word *Luft* (EN: air) is “nomen.Ort” (EN: noun.place), instead of “nomen.Substanz” (EN: noun.substance), which is listed as its second sense. *Familie* (EN: family) gets the semantic type “nomen.Kognition” (EN: noun.cognition), instead of “nomen.Gruppe” (EN: noun.group), which is listed as its second sense. For our experiment, this is also problematic, although it could be overcome by running a WSD task when tagging the texts, as mentioned earlier.

6 Conclusion and Future Work

In this paper, we have reported the results of a pilot experiment trying to retrieve automatically compound dictionaries from parallel corpora using WordNet. As we have seen, several problems arise, particularly regarding the usage of WordNets, but also other tools, such as the PoS taggers. The accuracy of the PoS taggers used here needs to be evaluated to identify sources of error that could be avoided. Eventually, a different PoS tagger may be needed.

More importantly, WordNet coverage for both languages is a problem. In our analysis of the full coverage pairs, we also realised that the matches of semantic types were not occurring as we expected. Two things may be done to overcome this problem. First, our semantic type matching needs to be revised. Second, a WSD task seems needed to ensure that the words in both languages are correctly tagged.

From this feasibility study, we may conclude this approach seems not to work for the task at hand. However, we have detected possible ways of improving the experiment setup which are worth future investigation. Additionally to the refinements already pointed out earlier, the use of supervised Machine Learning (ML) techniques may help to predict the Spanish phrase spans from the German compounds. Additional features such as the frequencies of apparition of certain words or spans of words in the Spanish corpus as well as possible PoS patterns could also help to filter the potential candidates better.

Acknowledgments

The research reported in this paper is supported by the People Programme (Marie Curie Actions) of the European Union’s Framework Programme (FP7/2007-2013) under REA grant agreements 238405 (project CLARA[1]) and 317471 (project EXPERT[2]).

[1] http://clara.b.uib.no/

[2] http://expert-itn.eu/

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by Machine of Complex Nominals: Getting It Right. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain.

Lars Bungum and Stephan Oepen. 2009. Automatic Translation of Norwegian Noun Compounds. In *Proceedings of the 13th Annual Conference of the EAMT*, pages 136–143, Barcelona, Spain, May.

Anthony Paul Cowie. 1998. *Phraseology: Theory, Analysis, and Applications: Theory, Analysis, and Applications*. Clarendon Press.

Christiane Fellbaum. 1998. *WordNet, An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language*, 19(4):479–496.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual Central Repository version 3.0. In *Proceedings of LREC'12*, Istanbul, Turkey, May.

Gregoy Grefenstette. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the 21st International Conference on Translating and the Computer*, London, United Kingdom. ASLIB.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid.

Su Nam Kim and Timothy Baldwin. 2013. Word sense and semantic relations in noun compounds. *ACM Transactions on Speech Language Processing*, 10(3), July.

Prashant Mathur and Soma Paul. 2009. Automatic Translation of Nominal Compounds from English to Hindi. In *Proceedings of the 7th International Conference on Natural Language Processing (ICON-2009)*, Hyderabad, India, December.

Hanne Moa. 2005. Compounds and other oddities in machine translation. In *15th Nordic Conference of Computational Linguistics*.

Rosamund Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford University Press.

Carla Parra Escartín. 2012. Design and compilation of a specialized Spanish-German parallel corpus. In *Proceedings of LREC'12*, Istanbul, Turkey, May.

Carla Parra Escartín. 2014. Chasing the Perfect Splitter: A Comparison of Different Compound Splitting Tools. In *Proceedings of LREC'14*, Reykjavik, Iceland, May.

Ulrike Rackow, Ido Dagan, and Ulrike Schwall. 1992. Automatic Translation of Noun Compounds. In *Proceedings of the 14th Conference on Computational Linguistics (COLING'92)*, Nantes, France.

Carlos Ramisch. 2012. *A generic and open framework for multiword expressions treatment: from acquisition to applications*. Ph.D. thesis, University of Grenoble (France) and Federal University of Rio Grande do Sul (Brazil), Grenoble, France, September.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword Expressions: A Pain in the Neck for NLP. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Takaaki Tanaka and Timothy Baldwin. 2003a. Noun-noun Compound Machine Translation: A Feasibility Study on Shallow Processing. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

Takaaki Tanaka and Timothy Baldwin. 2003b. Translation Selection for Japanese-English Noun-Noun Compounds. In *Proceedings of the Ninth Machine Translation Summit*, MT SUMMIT IX, New Orleans, USA.

Marion Weller and Ulrich Heid. 2012. Analyzing and Aligning German compound nouns. In *Proceedings of LREC'12*, Istanbul, Turkey, May.

Evaluation of Multiword Units Translation in Machine Translation: Another Pain in the Neck?

Johanna Monti
"L'Orientale" University
Naples, Italy
jmonti@unior.it

Amalia Todirascu
University of Strasbourg
Strasbourg, France
todiras@unistra.fr

Abstract

Recent studies have highlighted that the translation of Multiword Units (MWUs) by Machine Translation (MT) is still an open challenge, whatever the adopted approach (statistical, rule-based or example-based). The difficulties in translating automatically this recurrent, complex and varied lexical phenomenon originate from its lexical, syntactic, semantic, pragmatic and/or statistical but also translational idiomaticity. It is widely acknowledged that in order to achieve significant improvements in Machine Translation and translation technologies it is important to develop resources, which can be used both for Statistical Machine Translation (SMT) training and evaluation purposes. There is therefore the need to develop linguistic resources, mainly parallel corpora annotated with MWUs which can help improve the MT quality in particular as regards translation of MWUs in context and discontinuous MWUs. In this paper, we analyse the state of the art concerning MWU-aware MT evaluation metrics, the availability of both benchmarking resources and annotation guidelines and procedures.

1 Introduction

While MT quality evaluation has been a much-debated topic in MT since its inception, accurate MWU translation evaluation still poses a challenge, whatever the adopted MT approach (statistical, rule-based or example based). The main reason for this is that they display lexical, syntactic, semantic, pragmatic and/or statistical but also translational idiomaticity. Idioms, collocations, verb or nominal compounds, Named Entities or domain specific terms might all be considered as MWUs and in general both Statistical (SMT) or Rule-based Machine Translation (RBMT) fail to translate them correctly for different reasons, as highlighted by several recent contributions such as Barreiro et al. (2014), Monti (2012), and Ramisch et al. (2013) among others. MWU translation quality evaluation is not an easy task for two main reasons: lack of benchmarking resources and shared assessment methodologies and guidelines. MWU translation quality evaluation has so far not been discussed according to a shared methodological framework and, to the best of our knowledge, only very few MT quality evaluation metrics take issues related to MWU translation into account. For these reasons, to present there are only very few small-size corpora, containing aligned sentences representative of a specific type of MWU and a limited number of language pairs; these have generally been built to evaluate a specific MWU alignment tool or a specific MWU integration strategy in MT systems (Barreiro et al., 2014; Navlea and Todirascu, 2012; Ramisch et al., 2013; Weller et al., 2014). Annotated parallel corpora represent a very important resource since they are used in an MT development framework as training models from which SMT can extract and use the necessary information for the translation, or, in an evaluation framework, as benchmarking resources both to evaluate MT systems performances and help to improve their quality. To present, the availability of large data sets annotated with MWUs necessary for improving, on the one hand, MWU processing and translation in MT and on the other, quality estimation, is still very limited. As for many other NLP applications, the availability of annotated corpora represents the real bottleneck in relation to the technological and qualitative advancements in MWU processing and translation in MT. MWU annotation is a complex and time-consuming task since, given the current limitations in MWU identification by unsupervised methods, annotated resources are usually produced manually and in order to obtain large MWU annotated corpora, it is necessary to resort to a large number of experts in each specific language, which is not always an easy achievement. In addition, annotating MWUs in parallel corpora involves several problems:

- lack of agreement on the notion, the typologies, the boundaries of MWUs: this is a well-known problem and currently the COST Action IC1207 *PARSEME: PARSing and Multi-word Expressions. Towards linguistic precision and computational efficiency in natural language processing* is trying to solve it by means of a coordinated effort of multidisciplinary experts in different languages;
- translational asymmetries between languages. By translational asymmetries we mean the differences which occur between MWUs in the source language and their translations, like many-to-many (en. *to be in a position to* → it. *essere in grado di*) but also, many-to-one (en. *to set free* → it. *liberare*) and finally one-to-many correspondences (en. *overcooked* → it. *cotto troppo*);
- discontinuity: some MWUs admit insertions of external element, for instance verbal phrases such as in *take a [serious] haircut* or *count [Italy, Spain, and the Wednesday games winner] in.* This problem amplifies if we consider the non-isomorphism between languages in the annotation of discontinuous MWUs in parallel corpora.

In this contribution, we describe the state of the art concerning MWU-aware MT quality metrics in section 2, the availability of benchmarking resources in section 3 and finally the availability of annotation guidelines and procedures in section 4.

2 MWU-aware MT Quality Metrics

MT quality evaluation is a difficult task, mainly because there is no agreement on the quality parameters that have to be taken into account when assessing raw translation quality. The evaluation of MT outputs is twofold, since it is aimed at:

- estimating improvements or degradations in MT performance, on the one hand, mainly by means of unsupervised standard evaluation metrics, where MT outputs are compared with reference (human) translations. Well-known unsupervised metrics are BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) based on simple string similarity; METEOR based on shallow linguistic information such as lemmatisation, POS tagging and synonyms (Banerjee and Lavie, 2005); various other metrics including RTE (Pado et al., 2009) and ULC (Giménez and Màrquez, 2008) use deeper linguistic information such as semantic roles;
- evaluating MT usability in real work scenarios by Quality Estimation (QE) metrics, which do not have access to reference translations. The two most recent approaches in measuring MT quality are the TAUS Dynamic Quality Framework (DQF) and the QTLaunchPad Multidimensional Quality Metrics (MQM).

In both cases, the evaluation of MWU translation quality in MT outputs is not taken into account with the exception of a small number of MT evaluation metrics.

Babych and Hartley (2010) focus on a fine-grained MT evaluation, aiming at a more thorough error analysis of MWU translation. They adapt the BLEU metric to allow for the detection of systematic mistranslations of MWUs, and also to create a priority list of problematic issues. Two aligned parallel corpora serve as the basis for their experiments both with Rule-based (RBMT) and Statistical MT (SMT) systems. They show that their approach allows for the discovery of poorly translated MWUs both in the source and target language texts.

Stymne (2011) develops a fine-grained MT error analysis which includes concatenated definite noun phrases and compound nouns.

Salehi et al. (2015) present an attempt to integrate predicted compositionality scores of MWUs for English noun compounds into the TESLA machine translation evaluation metric.

3 Benchmarking Resources

Currently, MWU annotated benchmarking resources useful to translation quality evaluation are usually available for (i) specific MWU types (ii) specific language pairs (iii) evaluation of a specific MWU alignment tool or a specific MWU integration strategy in MT systems, (iv) assessment of the quality of different approaches to MWU in MT. In general evaluation corpora are manually built by carefully selecting sentences containing specific classes of MWUs to avoid data spareness. Thus, this time-consuming and difficult task sometimes produces resources which can very seldom be reused for other purposes. These resources are usually small-size parallel corpora containing either human translations collected from the Web or generated by MT systems (Google Translate, Bing, OpenLogos among others) and annotated manually. There are only very few instances of parallel corpora annotated with several types of MWUs and with different types of correspondences (many-to-one, one-to-many and many-to-many translations).

Ramisch et al. (2013) developed evaluation corpora annotated by several human annotators for verb compounds for English and French. They present two specific evaluation corpora: one corpus contains 17 highly fixed English idioms (eg. *kick the bucket*). It contains 10 sentences for each idiom and their manual translations. A second corpus, containing variable collocations (eg. *hold fire*) is also built, using sentences from the Internet and their human translations to Brazilian-Portuguese.

Barreiro et al. (2013) used a corpus of 150 English sentences extracted randomly from an existing corpus of sentences gathered from the news and the Internet. The corpus contains different multiwords, with an average of five MWUs per sentence. Each MWU under evaluation was annotated in the context of its sentence and classified according to a MWU taxonomy, developed in order to evaluate multiwords in any type of system, independently of the approach. The corpus was divided into three sets of 50 sentences each, and each set was then translated into French, Italian, and Portuguese respectively, using the OpenLogos and the Google Translate MT systems. The purpose of the study was not to compare and evaluate systems, but to assess and measure the quality of MWU translation independently of the two systems considered.

For French and Romanian Laporte (2014) presents a parallel reference corpus, composed of 1,000 pairs of sentences. The corpus is annotated with specific MWUs, such as Verb+noun constructions (idioms and collocations). MWUs were annotated by selecting the largest word sequences, both for continuous and discontinuous MWUs. MWUs in the source language are aligned either to simple words or MWUs in the target language.

Weller et al. (2014) built specific German-English corpora containing only compositional noun compounds, compositional verb compounds and a set of non-compositional compounds. All of the compounds occur in the training corpus. These annotated corpora were used to evaluate the output of a compound splitter and of this strategy with regard to the overall SMT system.

Barreiro et al. (2014) present a small parallel corpus containing 100 English Support Verb Constructions (SVC) that appear in sentences collected from the news and the Internet. The corpus was used to evaluate their translations into Italian, French, Portuguese, German and Spanish from English by a Rule-based machine translation (RBMT) system (OpenLogos) and a Statistical machine translation (SMT) system (Google Translate). The support verb constructions are classified by means of their syntactic structure and semantic behavior. The paper presents a qualitative analysis of the translation errors. The study aims to verify how machine translation (MT) systems translate fine-grained linguistic phenomena, and how well-equipped they are to produce high-quality translation.

Schottmüller and Nivre (2014) perform an evaluation of the translation quality of Verb-particle constructions (VPCs). They compare the results obtained from Google translate and Bing translate for German and English and offer a detailed analysis of translation errors. They have also made available a hand-made test corpus containing VPCs for the two languages.

Apart from manually built evaluation corpora, few parallel reference corpora are available. They usually contain a specific category of MWUs, such as light verb constructions. This is due to the various annotation problems, specific to each MWUs category (continuous vs discontinuous elements, MWUs delimitation, MWUs classification issues).

The SzegedParalell English-Hungarian parallel corpus Vincze (2012) constitutes the basis of the SzegedParalellFX, in which light verb constructions are manually annotated. Three novels, texts from magazines and language books and economic and legal texts were selected for annotation. Light verb constructions are annotated in both languages. The corpus has 14,261 sentence alignment units, which contain 1,370 occurrences of light verb constructions.

Rácz et al. (2014) describe 4FX, a quadrilingual (English-Spanish-German-Hungarian) parallel corpus annotated manually for light verb constructions. The 4FX corpus contains legislative texts from the JRC-Acquis Multilingual Parallel Corpus and contains 673 LVCs in English, 806 in German, 938 in Spanish and 1059 in Hungarian.

The CLUE corpus (Cross-Language Unit Elicitation alignments) [1] consists of a set of manual alignments of 400 parallel sentences from the Europarl corpora in four languages (Portuguese-English-Spanish-French), taking into consideration the following pairs: English-Spanish, English-French, English-Portuguese, Portuguese-Spanish. In order to establish the alignments between the different languages, for each considered language pair, two files are provided: the first one represents the word alignments, whereas the second represents the multiword units alignments.

This small number of instances of annotated corpora illustrate specific MWU and annotation guidelines are very seldom provided. For the development of future annotated corpora, methodological issues, annotation practices and guidelines should be proposed, from the past experiences.

4 Annotation Guidelines and Procedures

The lack of benchmarking resources annotated with MWUs is mainly due to the lack of shared assessment methodologies and guidelines and to the difficulties in annotating MWUs as highlighted in section 3. Thus, in addition to the scarcity of MWU annotated corpora, a further problem is represented by the fact that various annotation schemas or guidelines are adopted to annotate corpora. Annotation guidelines may include different MWU types (e.g. phrasal verbs, particle verbs, light verbs, compound nouns, named entities, idioms among many others) and the annotation schema may be more or less fine-grained (e.g. some annotation schema may consider only MWU types, whereas others include POS information, fixedness degree among others). Such differences can be attributed to the different purposes and applications for which they are developed. However, this situation represents a real obstacle to an effective reuse of existing annotated data. In addition, there are only very few papers which describe the annotation guidelines and procedures adopted and they usually address specific classes of MWUs (named entities, light verb constructions among others). For instance, Hwang et al. (2010) address the task of Proposition Bank (PropBank) annotation of light verb constructions. They have evaluated 3 different possible methods of annotation. The final method involves three steps: (1) manual identification of a light verb construction, (2) annotation based on the light verb construction's Frame File, and (3) a deterministic merging of the first two steps. They also discuss how in various languages the light verb constructions are identified and can be distinguished from the non-light verb word groupings. The developed multilingual schema for annotating LVCs takes into consideration the similarities and differences shared by the LVC as they appear in English, Arabic, Chinese, and Hindi.

Hendrickx et al. (2010) present a proposal for the annotation of multi-word units in a 1M corpus of contemporary Portuguese. Their aim is to create a resource that allows the study of MWUs in their context. The corpus is conceived as an additional resource next to the already existing MWU lexicon that was based on a much larger corpus of 50M words. The paper discusses the problematic cases for annotation and proposed solutions, focusing on the variational properties of MWUs.

Tutin et al. (2015) present an experiment of MWU annotation on the French part of a French-English bilingual corpus. Their aim is to achieve (i) a corpus-based and robust typology of MWUs; (ii) a basis for linguistic studies on MWUs, especially in relation to diverse textual genres; (iii) a corpus of evaluation for MT tasks, and especially SMT tasks.

Schneider et al. (2014) provide a useful description of their *comprehensive annotation approach*, in which all different types of MWUs are annotated in a 55K-word corpus of English web text. The guidelines adopted in the annotation take into account general issues (e.g., inflectional morphology; the spans of named entities; date/time/address/value expres-

[1] https://www.l2f.inesc-id.pt/~thomas/metashare/CLUE_narrative_description.pdf

sions;overlapping expressions), then briefly discuss 40 categories of constructions such as comparatives (*as X as Y*), age descriptions (*N years old*), complex prepositions (*out of, in front of*), discourse connectives (*to start off with*), and support verb constructions (*make a decision, perform surgery*). They also provide a detailed description of the annotation procedure, which was organised in three distinct phases: (a) individual annotation (a single annotator working alone); (b) joint annotation (collaborative work by two annotators who had already worked on the sentence independently); and (c) consensus annotation (by negotiation among three or more annotators, with discussion focused on refining the guidelines). They used a custom web interface for the annotation task. Creating general-purpose annotation guidelines is a difficult task, whereas projects such as Schneider et al. (2014) or Tutin et al. (2015) put forward valuable guidelines for English and for French. Such initiatives should be generalized for MWU translation evaluation.

5 Conclusions

We have outlined the state of the art concerning MWU translation evaluation in MT, and in particular of MWU-aware MT evaluation metrics, the availability both of benchmarking resources and annotation guidelines and procedures. Ongoing and future work includes recommendations for annotation guidelines which address the following issues: (i) the definition of various types of MWUs, both continuous and discontinuous, in order to give useful information for their identification and annotation, (ii) the selection of representative examples in the main European languages, (iii) a list of particularly challenging examples concerning alignment issues in the annotation of parallel corpora either in the form of simple parallel lists of MWUs or complete sentences.

Note

Johanna Monti is author of sections 1, 2 and Amalia Todirascu is author of section 4. Section 3, Abstract and Conclusions are in common.

Acknowledgments

We kindly acknowledge the anonymous reviewers for very useful feedback as well as the PARSEME European Cost Action (IC1207) for promoting this collaboration.

References

Bogdan Babych and Anthony Hartley. 2010. Automated error analysis for multiword expressions: using bleu-type scores for automatic discovery of potential translation errors. *Evaluation of Translation Technology*, 8:81.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, and Fernando Batista. 2013. When multiwords go bad in machine translation. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Machine Translation Summit XIV*, pages 26–33.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, Susanne Preuß, Kutz Arrieta, Wang Ling, Fernando Batista, and Isabel Trancoso. 2014. Linguistic evaluation of support verb constructions by openlogos and google translate. In *Proc. of LREC*, pages 35–40.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Jesús Giménez and Lluís Màrquez. 2008. A smorgasbord of features for automatic mt evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198. Association for Computational Linguistics.

Iris Hendrickx, Amália Mendes, and Sandra Antunes. 2010. Proposal for multi-word expression annotation in running text. In *Proceedings of the fourth Linguistic Annotation Workshop*, pages 152–156. Association for Computational Linguistics.

Elena-Mirabela Laporte. 2014. *La traduction automatique statistique factorisée: une application à la paire de langues français-roumain*. Ph.D. thesis, Université de Strasbourg.

Johanna Monti. 2012. *Multi-word unit processing in Machine Translation - Developing and using language resources for Multi-word unit* *processing in Machine Translation*. Ph.D. thesis, University of Salerno.

Mirabela Navlea and Amalia Todirascu. 2012. Using cognates to improve lexical alignment systems. In *Text, Speech and Dialogue*, pages 370–377. Springer.

S. Pado, M. Galley, D. Jurafsky, and C.D. Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of ACL- IJCNLP*.

Kishore Papineni, Salim Roukos, Ward Todd, and Wei-Jing Zhu. 2002. Bleu:a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318. Philadelphia.

Anita Rácz, István Nagy T., and Veronika Vincze. 2014. 4fx: Light verb constructions in a multilingual parallel corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Carlos Ramisch, Laurent Besacier, and Alexander Kobzar. 2013. How hard is it to automatically translate phrasal verbs from english to french? In *MT Summit 2013 Workshop on Multi-word Units in Machine Translation and Translation Technology*, page xx.

Bahar Salehi, Nitika Mathur, Paul Cook, and Timothy Baldwin. 2015. The impact of multiword expression compositionality on machine translation evaluation.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461. European Language Resources Association (ELRA), Reykjavik, Iceland.

Nina Schottmüller and Joakim Nivre. 2014. Issues in translating verb-particle constructions from german to english. In *10th Workshop on Multiword Expressions*, pages 124–131.

Sara Stymne. 2011. Pre-and postprocessing for statistical machine translation into germanic languages. In *Proceedings of the ACL 2011 Student Session*, pages 12–17. Association for Computational Linguistics.

Agns Tutin, Emmanuelle Esperana-Rodier, Manolo Iborra, and Justine Reverdy. 2015. Annotation of multiword expressions in french. In *Proceedings of Europhras 2015 (Europhras15)*. Malaga.

Veronika Vincze. 2012. Light verb constructions in the szegedparalellfx english–hungarian parallel corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Marion Weller, Fabienne Cap, Stefan Müller, Sabine Schulte im Walde, and Alexander Fraser. 2014. Distinguishing degrees of compositionality in compound splitting for statistical machine translation. In *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComAComA 2014)*, pages 81–90.

Statistical Measures to Characterise MWE Involving “mordre” in French or “bite” in English

Ismaïl El Maarouf
University of Wolverhampton
i.el-maarouf@wlv.ac.uk

Michael Oakes
University of Wolverhampton
m.oakes@wlv.ac.uk

Abstract

A large number of statistical measures exist which measure the collocational strength of Multi-Word Expressions (MWEs), particularly those which are characterised by two main words (Pecina, 2008). Such measures of collocational strength are useful for discovering new pairs of collocates in corpora. In this paper we will look at statistical measures of spread, flexibility, and diversity, which have not yet been tested for their ability to discover new collocates, but we have found useful for characterising MWEs containing collocates already found.

1 Introduction

This work is set in the context of the Corpus Pattern Analysis project (Hanks, 2004; Hanks, 2013), which aims to explore the relationship between word meaning and patterns of word use (in context). Word meanings are mapped onto phraseological patterns, rather than just being listed for words in isolation. CPA aims to provide well-founded resources for the analysis of meaning, particularly for Natural Language Processing and semantic processing, and its main output to date is the Pattern Dictionary of English Verbs (PDEV; `http://pdev.org.uk`; work in progress), which contains patterns for a large number of English verbs (more than 1,700 verbs).

Because CPA, and lexicography in general, is a time-consuming task, automating subtasks in the whole process of building a dictionary entry is highly desirable. The approach proposed in this paper uses statistical measures to speed up the discovery of specific types of verb-centred patterns, namely idioms. Particularly, it looks at how new statistical measures enable one to characterise different dimensions of idioms such as their spread, flexibility, and diversity. These measures are tested in a cross-lingual English/French perspective by focusing on idioms of the French verb “mordre” and the English verb “bite”, which are translations of one another in the literal sense.

Section 2 of this article presents CPA in more detail. Section 3 presents the new statistical measures and a worked out example. Section 4 analyses the results and section 5 concludes the paper.

2 Corpus Pattern Analysis

2.1 CPA

In CPA, lexicographers extract typical phraseological patterns from corpora by clustering corpus tokens (tagging them) according to the similarity of their context. Since we are concerned with the CPA of verbs, we will illustrate this technique with examples from verb analyses. The similarity between two corpus lines is evaluated in various ways, especially by performing syntactic and semantic analysis.

- Syntactic analysis involves the identification of the main verbal structures such as transitive/intransitive patterns, causative/inchoative alternations, argument/adjunct discrimination, idiomatic expressions, and phrasal verb uses.
- These structures are semantically analysed. For example, Semantic Types (ST; e.g. [[Human]], [[Building]], [[Event]]) are used to represent the prototypical properties shared by the collocates found in a specific pattern position.

Verb patterns can be described according to five main types of arguments: Subject, Object, Indirect Object, Complement, and Adverbial. Each argument can be characterised using a corpus-based apparatus of linguistic categories such as:

- Determiners, as in “*take place*” and “*take **his** place*”.

- Semantic types, as in "building [[**Machines**]] / [[**Relationship**]]".
- Contextual roles, as in '[[Human=**Film Director**]] / [[Human=**Sports Player**]] shoot'
- Lexical sets account for distinctions such as "reap the **whirlwind** / the **harvest**".

CPA has been applied only sparingly to nouns (Hanks, 2004; Hanks, 2012) and noun entries are quite different from verb entries. Some attempts at applying CPA to other languages have also been initiated, particularly Italian (Jezek et al., 2014) and Spanish (Renau and Battaner, 2012; Alonso Campo and Renau, 2013).

2.2 PDEV

For English, PDEV is the main electronic resource built using CPA. It contains more than 1,700 verbs for more than 4,600 patterns, as shown in Table 1[1].

Status	Entries	patterns
Completed verbs	1276	4601
Draft verbs	430	2154
Total verbs	1706	6755

Table 1: Number of entries and patterns in PDEV

Pattern set	entries	patterns
Phrasal Verbs	195	506
Idioms	200	456
Lexically Grounded	530	1268

Table 2: Idioms, phrasal verbs, lexically-grounded patterns (Complete)

Every time a lexicographer builds a new pattern, he/she indicates whether the pattern is (a) a phrasal verb or (b) an idiom. Moreover, as hinted at above, PDEV provides crucial information on other kinds of patterns which have some degree of fixedness, and ultimately, almost all patterns can be said to be fixed to some degree. A large number of PDEV patterns resort to lexical items: the previously mentioned example of "take place" belongs to this category, the Lexically-grounded patterns. The presence of a lexical item such as "place" is indeed a strong sign of fixedness. In Table 2, we can see that while idioms and phrasal verbs currently only constitute a small part of the patterns and entries, lexically-grounded patterns constitute about a quarter of the patterns.

3 Statistical Measures for MWE

3.1 Association measures

In psycholinguistics, "word association" means for example that people think of a term such as "nurse" more quickly after the stimulus of a related term such as "doctor". Church and Hanks (1990) redefined "word association" in terms of objective statistical measures designed to show whether a pair of words are found together in text more frequently than one would expect by chance. Pointwise Mutual Information (PMI) between word x and word y is given by Formula (1),

$$\boldsymbol{PMI(x,y)} = log_2 \frac{P(x,y)}{P(x).P(y)} \qquad (1)$$

where P(x,y) is the probability of the two words occurring in a common context (such as a span of 5 words, or in subject-object relation), while P(x) and P(y) are the probabilities of finding words x and y respectively anywhere in the corpus. PMI is positive if the two words tend to co-occur, 0 if they occur together as often as one would expect by chance, and less than 0 if they are in complementary distribution (Church and Hanks, 1990). PMI was used by Church and Hanks to examine the content word collocates of the verb "shower", which were found to include "abuse", "accolades", "affection", "applause", "arrows" and "attention". Human examination of these lists is needed to identify the "seed" members of categories with which the verb "shower" can occur, such as speech acts and physical objects, giving at least two senses of the verb (Hanks, 2012).

While PMI is useful for finding the strength of association between just two words, it can be extended to produce association measures for three words (Van de Cruys, 2011). Two variants suggested by Van de Cruys are Specific Correlation (SC) and Specific Interaction Information (SII), as shown in formula (2) and (3):

$$\boldsymbol{SC(x,y,z)} = log_2 \frac{P(x,y,z)}{P(x).P(y).P(z)} \qquad (2)$$

$$\boldsymbol{SII(x,y,z)} = log_2 \frac{P(x,y),P(y,z),P(x,z)}{P(x).P(y).P(z).P(x,y,z)} \qquad (3)$$

In a pilot experiment for the application of these measures, we found that highly scoring Subject-Verb-Object triples according to the SC measure were "Value added tax", "glazed UPVC window", "maximum branching ratio" and "stamped

[1] site accessed on May 2015

addressed envelope". However, the experiment found that many triples were low-frequency events which very often were coincidental recurrence of surface characteristics, such as capitalised text, and it seems that more meaningful results would be extracted from larger corpora (over 50 million words). To the best of our knowledge, little, if any, work has been made towards using these measures for MWE extraction.

3.2 Distance based statistical measures

Smadja (1993) recommends that collocations should not only be measured by their strength, such as by using the z-score, but also by their flexibility. We propose to implement this in the framework of text distance. Text distance is defined as the number of units between two units forming the boundaries of the expression of interest in a particular text. The units we use are words, but alternative kinds of units can be tested, such as characters. We propose to compute:

1. the **spread** of an expression, as the mean of the relative distances between two words forming the boundaries of an expression

$$\boldsymbol{\mu_{(X,Y)}} = \frac{1}{n}\sum_{i=1}^{n} dist(X_i, Y_i) \qquad (4)$$

2. the **flexibility** of an expression, which is the standard deviation of the relative distances between the two words

$$\boldsymbol{\sigma_{(X,Y)}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(dist(X_i, Y_i) - \mu_{(X,Y)})^2} \qquad (5)$$

High spread would indicate a flexible or semantic, rather than a rigid, lexical expression. In a study of David Wyllie's English translation of Kafka's Metamorphosis, Oakes (2012) found that "stuck fast" and "office assistant" had mean inter-word distances of 1 with a standard deviation of 0. This showed that, *in this particular text*, they were completely fixed expressions where the first word was always immediately followed by the second. Conversely, "collection" and "samples" had a mean distance of 2.5 with a standard deviation of 0.25. This word pair was a little more flexible, occurring both as "collection of samples" and "collection of textile samples". As a last example, "Mr. Samsa" had a mean distance of 1.17 and a standard deviation of 0.32. This is because it usually appeared as "Mr. Samsa" with no intervening words, but sometimes as "Mr. and Mrs. Samsa".

A third way of looking at the flexibility of an expression is by measuring the diversity of the forms found between the boundaries of this expression. A rigid expression, where all examples are identical in form and length, has very low diversity, while an expression which has many forms has much higher diversity. One measure of diversity, popular in ecological studies, is Shannon's diversity index, which is equivalent to entropy in information theory, and given by the formula

$$\boldsymbol{E_{(X,Y)}} = -\sum_{i=1}^{n} P_i log_2 P_i \qquad (6)$$

E is entropy, n is the number of *different* surface forms found for the expression, i refers to each surface form in turn, and P_i is the proportion of all surface forms made up of the surface form currently under consideration. The choice of logarithms to the base 2 ensures that the units of diversity are bits. The minimum value of diversity (when all the examples of a phrase or idiom are identical) is 0, while the maximum value (when all the examples occur in different forms) is the logarithm to the base 2 of the number of examples found. For standard deviation, the minimum value when all the examples are identical length is 0, and there is no theoretical upper limit.

3.3 Worked out example

In order to illustrate how values for each measure are computed we propose a worked out example based on a pair of words used as boundaries, "bite" and "dog", through a sample of 10 examples. The first thing to do is to compute the distance between those two words. First it is worth noting that we lump together alternative surface forms of the same boundary word, so we consider both "dogs" and "dog" as one word. Different decisions at this stage may lead to different results.

Figure 2 provides an example using signed distance (left or right): in the first example, dog is four words away to the left of bite. To compute the mean length, we recommend using the unsigned distance, but it is important to take into account the signed distance when computing standard deviation in order to capture word order differences. The unsigned text distances are therefore, in order of appearance of the examples, 4,4,3,2,4,1,3,2,1,2.

$$\boldsymbol{\mu'_{bite,dog}} = \frac{(-4)+(-4)+3+(-2)+4+(-1)+3+(-2)+(-1)+(-2)}{10}$$
$$= -0.6$$

$$\boldsymbol{\sigma_{bite,dog}} = \sqrt{\frac{(-4-(-0.6))^2+(-4-(-0.6))^2+(3-(-0.6))^2+(-2-(-0.6))^2+...+(-2-(-0.6))^2}{10}}$$
$$= \sqrt{\frac{76.4}{10}} = 2.76$$

Figure 1: Computation of standard deviation

The mean μ characterises the spread of an expression: "bite" and "dog" are 2.6 words apart.

$$\boldsymbol{\mu_{bite,dog}} = \frac{4+4+3+2+4+1+3+2+1+2}{10}$$
$$= 2.6$$

The standard deviation characterises the flexibility of an expression and is computed as illustrated in Figure 1. The score obtained for "bite" and "dog" is indicative of a high flexibility (2.76).

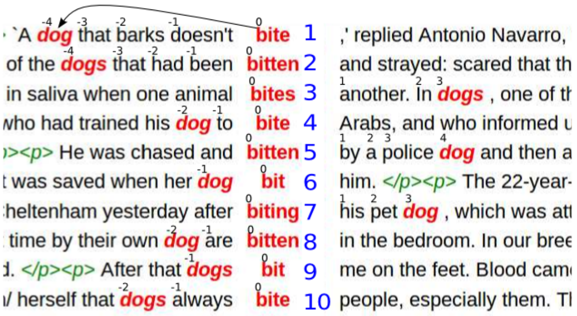
To compute Entropy, we must extract patterns of forms between boundaries. Again, either characters or words can be used as the basic unit, we here use words. The string between boundaries can also be characterised in various ways, and for our experiments, we use the surface forms of the words. A pattern is a full string between boundaries, and if there is no word, we consider it to be an instance of a null pattern. For $X = \{dog, dogs\}$, $Y = \{bite, bites, bit, bitten\}$, and $i = \{$*"that barks doesn't","that had been", "another. In", "to", "by a police","","his pet","are","always"*$\}$. P_i corresponds to the number of times the string is observed in the sample, divided by the total number of examples (in our case, 10). The entropy is computed as follows:

$$\boldsymbol{E_{bite,dog}} = -((\frac{1}{10}\log_2\frac{1}{10}) + (\frac{1}{10}\log_2\frac{1}{10}) + ...)$$
$$= 3.12$$

The entropy is quite high as there is no particular pattern that dominates in the sample: only the null pattern occurs twice, but the others, only once.

4 Results on bite and mordre

4.1 CPA



' `A dog that barks doesn't	bite	1	,' replied Antonio Navarro,
of the dogs that had been	bitten	2	and strayed: scared that th
in saliva when one animal	bites	3	another. In dogs , one of th
who had trained his dog to	bite	4	Arabs, and who informed u
></p><p> He was chased and	bitten	5	by a police dog and then a
t was saved when her dog	bit	6	him. </p><p> The 22-year-
heltenham yesterday after	biting	7	his pet dog , which was att
time by their own dog are	bitten	8	in the bedroom. In our bree
d. </p><p> After that dogs	bit	9	me on the feet. Blood cam
/ herself that dogs always	bite	10	people, especially them. Tl

Figure 2: 10 examples of "dog, bite"

A CPA analysis of the verb "bite" based on a sample of 500 lines of the British National Corpus found that it was used in 22 different patterns, 10 of them classified as idioms. The same process was applied to "mordre" on a sample of 500 lines from the Frtenten corpus (Jakubíček et al., 2013) and 16 patterns were found, 6 of them classified as idioms. The full list of patterns used with "bite" can be browsed on the PDEV website, and Table 3 lists the idioms found for "mordre". The rate of idioms per pattern is quite high for these two verbs, since, while about 12% of PDEV verbs have at least one idiom, an estimated 17% of this set of verbs have more than three idioms.

When comparing the meanings of these patterns, it was found that only the literal senses could be translated word for word without altering the meaning: these verbs share the meaning `[[Human 1 | Animal 1]] uses the teeth to cut into [[Animal 2 | Physical Object | Human 2]]`. Thus, French speakers would not use "mordre" but "piquer" to translate example (1) below, and, the verb "ronger" to translate example (2).

(1) The mosquitoes came up and bit me in the dark. *(Trans. Les moustiques sont venus et m'ont piqué dans le noir.)*

(2) A manicure is an effective way to stop biting your nails. *(Trans. La manucure est un moyen efficace pour arrêter de se ronger les ongles.)*

Num	Pattern and Implicature
4	**{[[Human]] | le poisson} mord ({à l'hamecon | a l'appat})** *[[Human]] takes the bait (= is lured to do something that has bad consequences)*
7	**[[Human]] mord {la vie à pleines dents}** *[[Human]] enjoys life to the full [literally, *bites life with full teeth]*
9	**[[Human]] se mord {les doigts}** *[[Human]] experiences a bitter time [literally, *bites his/her fingers]*
11	**[[Human 1]] fait mordre {la poussiere} {à [[Human]]}** *[[Human 1]] causes [[Human 2]] to bite the dust (= to die) or to lose a challenge [the latter sense only in French]*
12	**{le serpent} se mord {la queue}** [[Human]] is stuck in a [[State of affairs]] and cannot find a way out [literally, *the snake bites his own tail]
16	**[[Human]] ne mord pas [NO OBJ]** [[Human]] does not bite (= is harmless)

Table 3: List of idioms found in FrTenten for verb "mordre"

Collocate	Frequency	μ	σ	E
poisson	4	0	0	0
hameçon	4	3	0	0
appât	2	3	0	0
poussière	6	2	0	0
doigts	20	2.3	0.9	0.84

Table 4: Results for French idioms

collocate	Frequency	μ	σ	E
bullet	9	3	0	0
back	3	1	0	0
feed*	5	4	0	0
dust	10	2	0.09	0.15
bug	6	2	0.48	2.58

* including variants

Table 5: Results for English idioms

The situation is similar with idioms. Only one idiom can be translated word for word: "bite the dust". However, in French, it applies to two different contexts, either meaning "to die", or to "lose a challenge", so "mordre la poussière" can only be translated to English, word for word, if it means "die" in context. For all other idioms, translating word for word leads to a wrong interpretation. Hence, "biting one's fingers" has a non-literal meaning ("experiencing a bitter time") in French but not in English. Conversely, "biting one's tongue", has an idiomatic interpretation in English ("making a desperate effort not to say what is in his or her mind"), but not in French. These cross-lingual observations, identified with CPA, highlight the importance of dealing with MWEs in translation.

4.2 Statistical results

We applied the statistical measures to both French and English idioms identified in the sample CPA analysis, following recommendations in subsection 3.3. The results revealed that idioms come in a variety of forms, and have diverse properties on the scales of spread, flexibility, and diversity. Some examples are listed in Table 4 and Table 5.

In our experiments, we found that the idiom "bite the bullet" was maximally rigid, as it occurred all 9 times in exactly that form, with standard deviation and entropy both equal to 0. In contrast, the phrase "bitten by the ... bug" was extremely flexible, occurring all 6 times in different forms such as "bitten by the travel bug", "bitten by the London bug", and "bitten by the bug of the ocean floor". The standard deviation (0.48) was relatively small, reflecting that in all cases but one the variation consisted of the insertion of a single word, but the diversity index was its maximum value for 6 examples, log2(6) = 2.58.

In French, pattern 4 of "mordre" ("le poisson mord à hameçon/l'appât') comes in three different forms which have equivalent meaning ("to take the bait"), and have therefore been studied separately: the collocates "poisson", "hameçon", and "appât" were never found together. For each of them, independently of the spread (mean length) or its frequency, the idiom was always fixed (standard deviation = 0) and never took variants (Entropy = 0).

The idiom "[[Human]] se mord les doigts" usually occurred as "mordre les doigts", but sometimes as "mord encore les doigts" ("bites his fingers again"), "mordrait un peu trop souvent les doigts" ("bit his fingers a bit too often") and other variants. This gave a mean, standard deviation, and entropy of 2.3, 0.9, and 0.84 respectively.

The corresponding phrases "mordre la poussière" and "bite the dust" both have standard deviations and entropy close to 0, since, in both corpora, they allow very little variation.

Overall, the measures of standard deviation and entropy seem to coincide with intuitions about the rigidity of idioms, the scores being most of the time either equal to zero (maximally rigid and not diverse), or below 1, which still qualifies as low.

An important issue not touched upon in this paper, is the idiomaticity of an expression, that is the proportion of uses that do not have a literal interpretation. In a pilot study, we found that not all instances of a MWE defined as idiomatic, take an idiomatic reading in context. For example, we found that only 6 of the 24 expressions formed with the boundary words "kick" and "bucket", and in appropriate syntactic relation, had an idiomatic reading. Whether idiomaticity can be predicted from statistical measures is left to further experiments.

5 Conclusion and Perspectives

This paper has shown, through a cross-lingual Corpus Pattern Analysis of "bite" and "mordre", that MWEs are a major challenge in translation. Indeed, only a small portion of English patterns and collocations of "bite" can be translated word for word to "mordre" and preserve the meaning.

This paper suggests to use new statistical measures based on text distance to characterise the flexibility of a MWE, i.e. mean length, standard deviation of distances, and Entropy. These measures are intended to be used to speed up the discovery of CPA patterns for the Pattern Dictionary of English Verbs, but may also shed new light on other types of MWEs.

This article has attempted to evaluate whether these new measures are valid cross-linguistically, and found that to some extent, a low score in standard deviation and in Entropy indicate that the expression will tend to be a rigid idiom. However this paper has also found that idioms display a wide variety of characteristics, some of which can be captured by these statistical measures.

Future work in applying these measures cross-linguistically include exploring how to integrate them as features in Machine Translation systems in order to improve translation alignments.

Acknowledgments

This work was partly supported by a AHRC grant [DVC, AH/J005940/1, 2012-2015], a PARSEME grant [IC1207 COST Action], and EXPERT (People Programme (Marie Curie Actions) of the EU FP7 (REA grant agreement 317471)).

References

Araceli Alonso Campo and Irene Renau. 2013. Corpus pattern analysis in determining specialised uses of verbal lexical units. *Terminalia*, 7:26–33.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Patrick Hanks. 2004. Corpus pattern analysis. In *Proceedings of the XI EURALEX*, Lorient, France.

Patrick Hanks. 2012. How people use words to make meanings: Semantic types meet valencies. In A. Boulton and J. Thomas, editors, *Input, Process and Product: Developments in Teaching and Language Corpora*, pages 54–69. Brno, Czech Rep.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, MA.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The tenten corpus family. In *Proceedings of the International Conference on Corpus Linguistics*.

Elisabetta Jezek, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. T-pas; a resource of typed predicate argument structures for linguistic analysis and semantic processing. In *Proceedings of LREC*, Iceland.

Michael Philip Oakes. 2012. Describing a translational corpus. In Michael Philip Oakes and Meng Ji, editors, *Quantitative Methods in Corpus-Based Translation Studies*, "Studies in Linguistics" 51, pages 54–69. John Benjamins, Amsterdam.

Pavel Pecina. 2008. *Lexical Association Measures: Collocation Extraction*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic.

Irene Renau and Paz Battaner. 2012. Using cpa to represent spanish pronominal verbs in a learner's dictionary. In *Proceedings of the XV EURALEX*, Norway.

Frank Smadja. 1993. Retrieving collocation from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Tim Van de Cruys. 2011. Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, DiSCo '11, pages 16–20.

Aligning Verb+Noun Collocations to Improve a French - Romanian FSMT System

Amalia Todiraşcu
FDT, LiLPa,
Université de Strasbourg
22, rue René Descartes, BP 80010
67084 Strasbourg Cedex
todiras@unistra.fr

Mirabela Navlea
FDT, LiLPa,
Université de Strasbourg
22, rue René Descartes, BP 80010
67084 Strasbourg Cedex
mirabela_abe@yahoo.com

Abstract

In this paper we present two methods of integration of a specific class of multiword expressions, Verb+Noun collocations, into a French - Romanian factored statistical machine translation (FSMT) system. We adopt a static integration approach of Verb+Noun collocations: we identify and we transform them into single tokens in the training corpus, before the alignment step. Collocation identification is done by two methods: the application of a bilingual dictionary and the application of a hybrid extraction strategy from monolingual corpus. The evaluation shows that the dictionary-based method has slightly more influence to the overall performances of the FSMT than the hybrid method.

1 Introduction

Multiword expressions (MWEs) are composed of several words, with statistical, syntactic or semantic idiomaticity (Baldwin and Kim, 2010). MWEs include idiomatic expressions, verb or noun compounds, collocations, domain specific terms or Named Entities. Each class of MWEs is characterized by a set of specific properties. For example, idiomatic expressions are identified by high morphosyntactic fixedness and non-compositional sense. Verbal collocations have variable syntactic structures (accepting modifiers) and a more compositional sense.

MWEs represent a "pain in the neck" (Sag and Wasow, 2002:1) for machine translation (MT) systems. MWEs require specific integration strategies into MT systems. The main translation problems of MWEs are due to their specific properties:

- **non-compositionality**. *mettre de l'huile sur le feu* (French) "to put fuel on the fire" vs. *a pune paie pe foc* (Romanian) "to put fuel on the fire". In this example, MWEs express the same idea, but a direct word-to-word translation is not possible.

- **strong lexical preferences**. *donner une conférence* (French) "give a talk" vs. *a ține o conferință* (Romanian) "give a conference". In this case, the verb *a ține* "to hold" is not synonym to the verb *donner* "to give", in other contexts. The synonym *tenir* "to hold" is not accepted in a construction such as **tenir une conférence*.

- **many-to-one translations**. A MWE might be translated by a single word in the target language: *jeter l'ancre* (French) "to anchor" vs. *a se stabili* (Romanian) "to settle".

Statistical machine translation (SMT) systems generally fail to handle correctly these problems, due to data sparseness, but also due to the fact that MWEs are not properly detected in automatic manner. If FSMT systems (Koehn et al, 2007) use linguistic information to avoid data sparseness, specific strategies are proposed to handle MWEs in SMT and in FSMT systems. Several SMT methods adopt a static integration of MWEs, by transforming them as single tokens in the training corpora, before the alignment step (Ramisch et al., 2013). Other strategies replace specific classes of MWEs such as phrasal verbs with paraphrases (Barreiro, 2009) or with their litteral meaning (Salton et al., 2014) or by using external resources (bilingual dictionaries or term glossaries) (Kordoni and Simova, 2014; Ren et al, 2009). On the one hand, static strategies introduce MWEs as single units into the training data and the overall MT procedure is the same as a classical SMT (Pal et al., 2011). On the other hand, dynamic integration approaches include MWEs by completing word alignment (Okita and Way, 2011; Wu et al., 2008) or by completing training data with MWEs and their translation equivalents integrated directly into phrase tables (Kordoni and Simova, 2014; Bouamor et al, 2012).

In this project, we adopt a static collocation integration strategy (Ramisch et al., 2013) into a factored statistical machine translation (FSMT) system for French and Romanian (Navlea, 2014). We focus on a specific class of MWEs: Verb+Noun collocations (Gledhill, 2007). We identify Verb+Noun collocations by two methods. The first method exploits an external bilingual dictionary (Todiraşcu et al., 2008). The second method uses a hybrid strategy to extract Verb+Noun collocations from monolingual corpus (Todiraşcu et al., 2009).

We present the collocation definition adopted in our project in the next section. In section 3, we present the architecture of the FSMT and the various modules detecting collocations. Section 4 presents the training and the test corpus, while section 5 and 6 present the dictionary-based collocation integration method and the hybrid extraction method, respectively. The section 7 presents a comparison of the two methods.

2 Verb+Noun Collocations

Collocations represent multiword expressions, sometimes discontinous, with specific syntactic and semantic behavior (Gledhill, 2007). Collocations are identified through three criteria (Gledhill, 2007; Gledhill and Todirascu, 2008): frequency, syntactic dependencies and semantic properties. Collocations are frequent word associations (Sinclair, 1991), mapping several syntactic patterns (Verb+Noun, Noun+Noun, Adverb+Adjective) (Hausmann, 2004). They present an important degree of non-compositionality.

The Verb+Noun collocations class is domain-independent. These collocations are characterized by a syntactic dependency between the verb and the noun. The noun generally specifies the range (Halliday, 1985) of the process expressed by the verb (Gledhill, 2007). We study two classes of Verb+Noun collocations (Gledhill, 2007), such as:

a) **Complex predicators**. This class of collocations is characterized by high fixedness (strong preferences for morpho-syntactic and syntactic properties). Indeed, the noun accepts only some specific types of determiners, some specific values for the number, gender or case. The verb and the noun form together the predicate. The sense of these collocations is non-compositional and they generally express a relational process. This class includes idiomatic expressions (*mettre de l'huile sur le feu* "to put fuel on the fire"), but also Verb+Noun collocations with high mophosyntactic fixedness (*jeter les bases* "to lay the bases");

b) **Complex predicates**. This class has more variable morphosyntactic and syntactic properties, such as accepting a large set of determiners. They are more flexible in accepting modifiers or passive forms. The sense of these collocations is more compositional and they generally express a mental process (*prendre des décisions* "take decisions", *faire une erreur* "to make a mistake"). Syntactically, the noun is the direct object of the verb.

Handling Verb+Noun collocations, as other categories of MWEs in SMT is a difficult task. Thus several strategies of their integration are domain-dependent. As we focus here on a specific class of collocations, which is domain-independent, we adopt a method of static integration of collocations into a FSMT system. The architecture of this system is described in the next section.

3 The Architecture of the FSMT

Our French - Romanian FSMT system is based on the open-source MOSES decoder (Koehn et al., 2007). The baseline factored translation model (Navlea, 2014) is trained by using linguistic factors such as lemmas and morphosyntactic properties. This baseline system uses the *grow-diag-final* training heuristic (Koehn et al., 2003), which exploits the intersection, but also the union of bidirectional lexical alignments. The lexical alignment of the training parallel corpus is performed by using GIZA++ statistical aligner (Och and Ney, 2003). In the French -> Romanian translation direction, existing Romanian language models (Tufiș et al., 2013) are exploited, while in the opposite translation direction, our own French language models are built by using SRILM application (Stolcke, 2002). These models are based on the law corpus JRC-Acquis (Steinberger et al., 2006). They are developed on surface word forms or on different linguistic factors such as lemmas and morphosyntactic tags.

In this paper, we study the influence of the MWEs detection and alignment on the factored baseline system. Thus, the architecture of the final system includes two different modules of MWEs integration:

- the first module applies a bilingual collocation dictionary (Todiraşcu et al., 2008) to detect MWEs in the training corpus and then transforming them into single tokens (Ramisch et al., 2013);
- the second module exploits a hybrid MWEs extraction method (Todiraşcu et al., 2009) to identify MWEs from monolingual corpus and then transforming them into single tokens in the training corpus (Ramisch et al., 2013);

For both methods, collocation alignment is also realized by GIZA++ (Och and Ney, 2003).

We give in the Figure 1 the architecture of the French - Romanian FSMT system integrating the two modules of collocation identification.

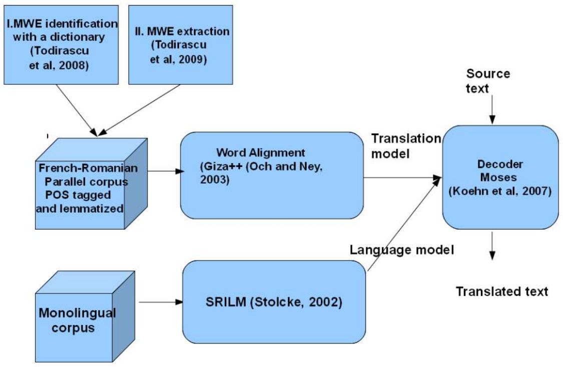

I.MWE identification with a dictionary (Todirascu et al, 2008)
II. MWE extraction (Todirascu et al, 2009)
Source text
French-Romanian Parallel corpus POS tagged and lemmatized
Word Alignment (Giza++ (Och and Ney, 2003)
Translation model
Decoder Moses (Koehn et al, 2007)
Language model
Translated text
Monolingual corpus
SRILM (Stolcke, 2002)

Figure 1. The architecture of the French - Romanian FSMT system

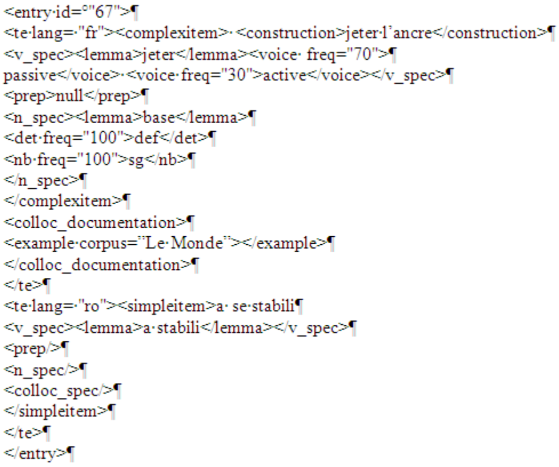
4 The Corpus

For our experiments, we use a legal training parallel corpus containing about 1.5 million of tokens per language (64.918 bilingual parallel sentences). In order to test, but also to optimize the FSMT system, two separate small parallel corpora composed each of 300 bilingual parallel sentences are kept. All these corpora are extracted from the *DGT-TM* translation memory (Steinberger et al., 2012). These corpora are preprocessed (tokenization, lemmatization, tagging, and chunking) with TTL tagger available for Romanian (Ion, 2007) and for French (Todiraşcu et al., 2011). Each token is associated with linguistic information such as lemmas (followed by the two first characters of the morphosyntactic tag to morphologically disambiguate the lemmas), part-of-speech, and morphosyntactic tag. TTL uses the MSD (morphosyntactic descriptors) proposed in the MULTEXT project (Ide and Veronis, 1994), for French, and in the MULTEXT-East project (Erjavec, 2004) for Romanian. Aditionnaly, the Romanian chunker identifies several domain-specific terms (such as Noun+Noun terms), while the French chunker does not have this feature.

Initially, in these corpora the Verb+Noun collocations are not identified. We detect them for our experiments and we transform them into single tokens in the corpora.

5 The Dictionary-based Integration Method

In the first method, we use a small bilingual collocation dictionary (Todiraşcu et al., 2008) to identify the collocations and to transform them into single units. The dictionary contains 250 bilingual entries. Each entry represents the translation equivalent and rich information about the morphosyntactic properties of the collocations. The properties of the verb (voice, auxiliary, lexical preferences) and of the noun (preferences for a specific number, case, gender, or determiner) are detailed for each collocation (see Figure 2).



```
<entry·id=°"67">¶
<te·lang=·"fr"><complexitem>·<construction>jeter·l'ancre</construction>¶
<v_spec><lemma>jeter</lemma><voice· freq="70">¶
passive</voice>·<voice·freq="30">active</voice></v_spec>¶
<prep>null</prep>¶
<n_spec><lemma>base</lemma>¶
<det·freq="100">def</det>¶
<nb·freq="100">sg</nb>¶
</n_spec>¶
</complexitem>¶
<colloc_documentation>¶
<example·corpus="Le·Monde"></example>¶
</colloc_documentation>¶
</te>¶
<te·lang=·"ro"><simpleitem>a· se·stabili¶
<v_spec><lemma>a·stabili</lemma></v_spec>¶
<prep/>¶
<n_spec/>¶
<colloc_spec/>¶
</simpleitem>¶
</te>¶
</entry>¶
```

Figure 2. Example of a bilingual entry

To identify collocations in the source corpus, we search for the pairs of verbs and of nouns found in the dictionary. Then, according to the various specific collocation properties (preferences for a specific class of determiners, for a specific preposition) we identify all the words that should be included in the collocation. For each source sentence containing the collocation found in the dictionary, we search the translation equivalent and its specific properties. The source collocation and its equivalents are transformed into single tokens.

6 The Hybrid Extraction Method

The second method of collocations integration into the FSMT system uses a hybrid extraction strategy to obtain candidates from monolingual corpus. Then, the extracted candidates are transformed into single tokens in the training corpora, before launching the alignment process.

This hybrid method aims to identify Verb+Noun collocation candidates from a monolingual corpus. For French, we use a corpus of approx 1 million tokens composed of texts extracted from JRC-Acquis (Steinberger et al., 2006), from medical corpora and from newspapers corpora (LeMonde, Est Républicain). For Romanian, we use a corpus of approx. 0.5 milions of tokens composed of law texts (from JRC-Acquis) and newspapers texts. We apply the hybrid identification method (Todirascu et al., 2009) combining the frequency criteria and the morphosyntactic information about each class of Verb+Noun collocations. This module extracts frequent word associations, with the Log-likelihood (LL) (Dunning, 1990) greater than 9. In our experiments, we extract pairs of verb and nouns, occurring in a window of 11 words. This statistical extraction has as results some irrelevant candidates.

The linguistic information (POS tags, lemmas) are used to filter out some of these irrelevant candidates, but also to identify possible complex predicators or predicate candidates. Indeed, complex predicators are characterized by high morphosyntactic fixedness. Linguistic filters identify them by checking all the occurrences and their contexts. If the candidate has high preference (more than 85 % of the contexts) for some specific morphosyntactic configurations, then this could be considered as a complex predicator candidate.

A filter identifying Romanian complex predicator (*a aduce atingere* ‘to make damage’, a da naştere ‘to give birth’) exploits the preference for the zero determiner and for singular. The filter is given in CQP format:

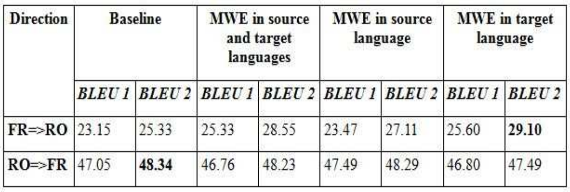
`[pos=”Vm.”][pos=”Nc.s-n”]` where Vm. means main verb, Nc.s-n means Nc common noun, s- singular, n – without determiner.

Some filters use heuristics to identify irrelevant candidates (for example, if the verb and the noun are separated by several prepositional groups or conjunctions, then the verb and the noun do not form a collocation). This category of filters identify 40.15% of irrelevant candidates for Romanian and 39.43% for French. For our experiments we selected only the candidates with LL > 5000 to be transformed into single tokens.

7 Comparing the Methods

In this paper, we evaluate two strategies to integrate MWEs such as Verb+Noun collocations (Gledhill, 2007) into a baseline French - Romanian FSMT system (Navlea, 2014). These strategies are the application of a bilingual dictionary (Todiraşcu et al., 2008), but also of a hybrid MWEs extractor from monolingual corpus (Todiraşcu et al., 2009) to identify Verb+Noun collocations in the training corpora and then to align them (Ramisch et al., 2013). In order to evaluate the systems, we use the BLEU (Bilingual Evaluation Understudy) score (Papineni et al., 2002) as fol-lows: BLEU 1 - the score before system tunning, BLEU 2 - the score after system tunning. The optimization step is performed by using MERT application (Bertoldi et al., 2009).

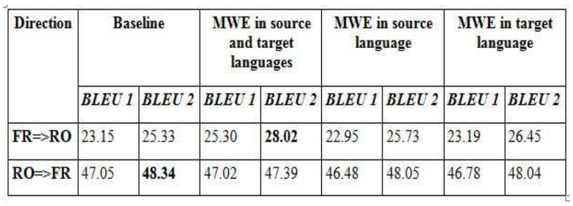
We apply the first identification method to the training corpus, for both languages, only in the source language and only in the target language (see Table 1). We run the FSMT system into both directions (French=>Romanian and Romanian=>French). The application of the dictionary is slightly better for French=>Romanian translation direction, improving the BLEU 2 score in all the configurations. The best score is obtained when Verb+Noun collocations are included into target language (3.77 points of improvement). In the other direction, the BLEU 2 score obtained for the baseline is the highest (48.34), but the configurations of Verb+Noun collocation identification in both languages (48.23) or in source language (48.29) are quite similar with the baseline.



Direction	Baseline		MWE in source and target languages		MWE in source language		MWE in target language	
	BLEU 1	*BLEU 2*	*BLEU 1*	*BLEU 2*	*BLEU 1*	*BLEU 2*	*BLEU 1*	*BLEU 2*
FR=>RO	23.15	25.33	25.33	28.55	23.47	27.11	25.60	**29.10**
RO=>FR	47.05	**48.34**	46.76	48.23	47.49	48.29	46.80	47.49

Table 1. Dictionary-based method for collocation identification and the results of the FSMT system

The second set of experiments is done with the list of Verb+Noun collocations obtained with the hybrid method. We also compare three various configurations (by detecting collocations in both languages, in the source language and in the target language) (see Table 2). The best improvement of the BLEU score (2.69 points) is again obtained for the French=>Romanian direction when collocations are detected in both languages. However, in the opposite direction, the baseline obtain the best score.



Direction	Baseline		MWE in source and target languages		MWE in source language		MWE in target language	
	BLEU 1	*BLEU 2*	*BLEU 1*	*BLEU 2*	*BLEU 1*	*BLEU 2*	*BLEU 1*	*BLEU 2*
FR=>RO	23.15	25.33	25.30	**28.02**	22.95	25.73	23.19	26.45
RO=>FR	47.05	**48.34**	47.02	47.39	46.48	48.05	46.78	48.04

Table 2. Hybrid MWE extraction method for collocation identification and the results of the FSMT system

The BLEU scores are better for Romanian=> French translation direction. This is due to the fact that Romanian has rich morphology, helping to generate the right surface form. Romanian corpus contains a set of domain-specific terms, while for French these elements are not identified.

The dictionary obtains slightly better improvement (1 point) at least for the French=>Romanian direction, rather than the hybrid extraction method. However, the dictionary is quite small (250 entries) and the coverage between the dictionary and the test corpus is small as well : 14 French collocations have 27 occurrences in the French test corpus, while 15 Romanian collocations were found 34 times in the Romanian test corpus. Further experiments should be done with other test corpus (containing more collocations) and with a larger dictionary.

8 Conclusion and Further Work

We present two static integration methods of a specific class of MWEs: Verb+Noun collocations. We identify the Verb+Noun collocations before the alignment process and we transform them into single units. We compare these two methods of collocation identification: a dictionary-based strategy and a hybrid extraction method. The dictionary obtained slightly better results (1 point) rather than the hybrid method, at least for French=> Romanian translation direction. Further experiments should be done with a larger dictionary to validate this conclusion. Further work consists of a classification of collocation translation errors in order to explain the various differences between the configurations.

In further experiments we will compare the results of these two methods with the results provided by another static strategy of mapping collocations via an extended dictionary (Navlea, 2014) and dynamic integration with the help of a specific alignment algorithm.

Reference

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In N. Indurkhya, & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (pp. 267-292), Second edition. Boca Raton (USA, FL): CRC Press, Taylor and Francis Group.

Anabela Barreiro. 2009. *Make it simple with paraphrases: Automated paraphrasing for authoring aids and machine translation,* Ph.D.Thesis, University of Lisbon

Nicola Bertoldi, Barry Haddow and Jean-Baptiste Fouet. 2009. Improved Minimum Error Rate Training in Moses. *Prague Bulletin of Mathematical Linguistics (PBML)*, 91, 7-16.

Dhouha Bouamor, Nasredine Semmar and Pierre Zweigenbaum. 2012. Identifying bilingual multiword expressions for statistical machine translation. In *LREC 2012, Eigth International Conference on Language Resources and Evaluation*, pages 674-679, Istanbul, Turkey, 2012. ELRA.

Ted, Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), 61-74.

Tomaz Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Paris: ELRA, pp. 1535-1538

Christopher Gledhill. 2007. La portée : seul dénominateur commun dans les constructions verbo-nominales. In Frath, P., Pauchard, J., & Gledhill, C. (Eds.), Actes du 1er colloque, Res per nomen, pour une linguistique de la dénomination, de la référence et de l’usage (pp. 113-125), Université de Reims-Champagne-Ardenne.

Christopher Gledhill and Amalia Todiraşcu. 2008. Collocations en contexte : extraction et analyse contrastive. Texte et corpus, 3, Actes des Journées de la Linguistique de Corpus 2007 (pp. 137-148).

Michael Halliday. 1985. *Introduction to Functional Grammar*. London: Edward Arnold.

Franz Joseph Hausmann. 2004. Was sind eigentlich Kollokationen?. In K. Steyer (Ed.), *Wortverbindungenmehr oder weniger fest*. Institut fur Deutsche Sprache Jahrbuch, pp. 309-334

Nancy Ide and Jean Véronis. 1994. Multext (multilingual tools and corpora). In *Proceedings of the 15th CoLing*. Kyoto (Japon), pp. 90-96

Radu Ion. 2007. *Metode de dezambiguizare semantică automată. Aplicații pentru limbile engleză şi română (in Romanian) Semantic desambiguation methods. Applications for English and Romanian languages*. Ph.D.Thesis. Bucharest (Roumania): Romanian Academy.

Philip Koehn, Franz Och and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Ed-monton (Canada). Stroudsburg (USA, PA): Asso-ciation for Computational Linguistics, pp. 48-54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico; Nicola Bertoldi, Brooke Cowan, Wade Shen; Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin and Evan Herbst. 2007. Moses : Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*. Prague, pp. 177-180.

Valia Kordoni and Iliana Simova. 2014. Multiword Expressions in Machine Translation, *Proceedings of LREC 2014*, Reykjavík, Iceland

Patrick Lambert and Rafael Banchs. 2006. Grouping multi-word expressions according to Part-Of-Speech in statistical machine translation. In *Proceedings of the EACL Workshop on Multi-word expressions in a multilingual context*. Trento, Italy, pp. 9-16.

Franz J. Och and Herman Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29(1):19-51.

Tsuyoshi Okita and Andy Way. 2011. MWE-sensitive Word Aligner in Factored Translation Model. *Machine Translation and Morphologically-rich Languages: Research Workshop of the Israel Science Foundation*, Israel, Jan, 2011. pp.16-17.

Mirabela Navlea. 2014. *La traduction automatique statistique factorisée, une application à la paire de langues français - roumain*, Ph.D.Thesis, Université de Strasbourg, France, June 2014.

Santanu Pal, Tanmoy Chakraborty, and Sivaji Bandyopadhyay. 2011. Handling multiword expressions in phrase-based statistical machine translation. In *Proceedings of the 13th Machine Translation Summit*, MT Summit 2011, pp. 215–224.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002.. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual meeting of the Association for Computational Linguistics (ACL)*, Philadelphia (USA, PE). Stroudsburg (USA, PA): Association for Computational Linguistics, pp. 311-318

Carlos Ramisch, Laurent Besacier and Alexander Kobzar. 2013. How hard is it to automatically translate phrasal verbs from English to French?. In J. Monti, R. Mitkov, G.. Corpas Pastor, V. Seretan (Eds.), *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology*, Nice (France), pp. 53-61.

Zhixiang Ren, Yajuan Luo, Jie Cao, Qun Liu and Yun Huang. 2009. Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions. In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, pp. 47-54.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword Expressions:A Pain in the Neck for NLP. A. Gelbuch (Ed.) *Proceedings of CICLing 2002*, LNCS 2276, pp. 1-15.

John Sinclair. 1991. *Corpus, Concordance, Collocation*, Oxford University Press.

Ralph Steinberger, Andreas Eisele, Klocek, S., Pilos, S., and Schlüter, P. 2012. DGT-TM: A freely Available Translation Memory in 22 Languages. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*. Istanbul (Turquie): ELRA, pp. 454-459.

Andreas Stolcke. 2002. SRILM -- An Extensible Language Modeling Toolkit. *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, Denver, pp. 901-904.

Amalia Todiraşcu, Radu Ion, Mirabela Navlea and Laurence Longo (2011). French text preprocessing with TTL. In *Proceedings of the Romanian Academy*, Series A, 12(2):151-158.

Amalia Todiraşcu, Christopher Gledhill and Dan Stefanescu. 2009. Extracting Collocations in Contexts. In Z. Vetulani and H. Uszkoreit (Eds.), *Responding to Information Society Challenges: New Advances in Human Language Technologies, LNAI 5603*. Berlin Heidelberg: Springer-Verlag, pp. 336-349.

Amalia Todiraşcu, Ulrich Heid, Dan Ştefanescu, Dan Tufiş, Christopher Gledhill, Marion Weller, François Rousselot, 2008. Vers un dictionnaire de collocations multilingue. In Xavier Blanco Escoda, Marie-Claude L'Homme et Marc Van Campenhoudt (éd) *Special Issue on "Lexique, dictionnaire et connaissance dans une société multilingue, "CAHIERS DE LINGUISTIQUE, Revue de sociolinguistique et de sociologie de la langue française,* Éditions Modulaires Européennes, vol. 33/1 : 161-186.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester. pp. 993–1000

Multiword Expressions in Machine Translation: The Case of German Compounds

Maria Ivanova **Eric Wehrli** **Luka Nerima**

University of Geneva
{maria.ivanova|eric.wehrli|luka.nerima}@unige.ch

Abstract

This paper presents an on-going research on compound generation for the German language, to be integrated into a speech-to-speech machine translation system with English, French and Italian as a source language and German as a target language. Currently, we restrict ourselves to nominal compounds and furthermore only to compounds of type noun-noun. We developed a module, which combines different observations from theoretical linguistics in order to find the correct connecting elements in German noun compounds. The module outperforms the baseline system by 16%.

1 Introduction

The importance of multiword expressions in translation has been long recognized. Their proper identification and well-formed generation are fundamental requirements for high-quality translation. German compounds constitute a particularly important case, due to their high frequency, variety and high productivity, ruling out any hope to simply list them all in a dictionary (Parra Escartín et al., 2014).

This paper presents an on-going research on compound generation for the German language, to be integrated into a speech-to-speech machine translation system with English, French and Italian as a source language, and German as a target language[1]. Currently we are working on the English-to-German translation pair. We restrict ourselves to nominal compounds and furthermore only to compounds of type noun-noun (e.g. *apple tree → Apfelbaum, toothbrush → Zahnbürste*).

Two distinct issues arise with respect to German compound generation: when to generate a compound, and how precisely to form it. The first one concerns the determination of what source structure should trigger the generation of a compound (e.g. noun-noun as in *chocolate cake*, noun-prep-noun as in *fall in population*, etc.), and in which case (e.g., not all noun-noun structures are compounds). The second issue concerns the precise way of combining the constituents, that is to say what connecting morpheme (or connecting element) should be used to glue the two nouns. German noun compounds are built either by glueing the constituents of the compounds directly (zero morpheme), or by using a connecting element in-between, such as *-s* or *-en*, to merge the constituents. Determining the connecting morpheme of a compound is a non-trivial task (Žepić, 1970; Ortner et al., 1991; Fuhrhop, 1996; Fuhrhop, 1998).

The main focus of this paper is on the second issue. We have developed a compound generation module using lexical and linguistic information (e.g., inflectional paradigm, morphology, gender). First we analyze the compounds in order to retrieve the basic lexemes and then recompose using the German generation module, thus achieving a true German-to-German translation task.

2 Connecting Elements and Theoretical Linguistics

In this work we focus only on compounds of type noun-noun. According to the official German spelling, the constituents of noun compounds in German are always written together, e.g. *Apfelbaum*, and sometimes they can be connected with a hyphen, e.g. *Kaffee-Ersatz*[2] (Rechtschreibung, 2006). When the constituents are combined, a connecting element, e.g., *-s, -en, -er*, can be appended to the first compound constituent. Sometimes the first constituent might be truncated, e.g., *Hilfe* but *Hilf-**s**-arbeit*[3]. Often the connecting ele-

[1] The work is part of the SIWIS project (Garner, 2014)

[2] *coffee substitute*

[3] *help* but *unskilled labour*

ments in German noun compounds coincide with inflectional suffixes, e.g. *Staat*-**s**-*vertrag*, *Student*-**en**-*haus*[4]. In other cases, however, they do not: *Gesundheit*-**s**-*amt*, *Hahn*-**en**-*feder*[5]. And yet in other cases, more than one connecting element might be used with the same first constituent, e.g., *Tag*-**e**-*buch*, *Tag*-**es**-*licht*[6] or *Kind*-**er**-*garten*, *Kind*-**s**-*kopf*, *Kind*-**es**-*beine*[7].

Despite this “arbitrariness”, native speakers use connecting elements in a rather uniform way. It is believed that there exist rules and distributional restrictions, which native speakers use subconsciously, and that the *feeling* for the right building of compounds is the *knowledge* of the rules, the restrictions and the possibilities, which exist in the language (Žepić, 1970; Fuhrhop, 1996; Fuhrhop, 1998). Having this in mind, the problem of assigning the correct connecting element in a compound becomes the discovery of these rules and restrictions.

2.1 Morphology of German Noun Compounds

In theoretical linguistics a German compound is formally defined as $A + V + B$, where A and B are respectively the first and second compound constituents, and V is the connecting morpheme (Žepić, 1970). V is the morpheme that remains after the constituent A has been shortened to its base form (nominative singular). Compounds can be built recursively, i.e., A and B can be compounds themselves.

It has been observed that the morphology of noun compounds largely depends on the grammatical features of the first compound constituent, especially on its inflectional paradigm[8], phonetic structure, gender, scope, and partly on its relation with the second compound constituent (Ortner et al., 1991). In this sense, the connecting morpheme is a part of constituent A and a noun compound can be defined more clearly as $(A+V)+B$ (Žepić, 1970).

Table 1 shows different connecting elements and example contexts in which they appear, as presented in (Žepić, 1970)[9]. Restrictions in the distribution of these elements have been discovered by conducting empirical research. Some of these restrictions hold deterministically, others are only partially deterministic. For example, it has been observed (Žepić, 1970; Ortner et al., 1991) that a noun with the suffix *-ung*, e.g. *Versicherung*[10], can regularly take only the *-s* connecting element, e.g. *Versicherung-s-vertrag*[11]. Another example involves the inflectional class of A: Žepić (1970) showed that the classification of first compound constituents into their inflectional classes reduces the number of connecting elements for A from the 9 possible morphemes (see Table 1) to a choice of only 1 to 4 morphemes[12]. A thorough description of many observed distributional rules and restrictions can be found in literature (e.g. (Žepić, 1970; Ortner et al., 1991; Fuhrhop, 1996; Fuhrhop, 1998))[13].

Conn. morphemes	**Examples**
A + **Ø** + B	*Dampf-maschine*[14]
A + **(e)s** + B	*Jahr***es***-zeit, Antritt***s***rede*[15]
A + **(e)n** + B	*Elektron***en***-röhre*[16]
A + **er** + B	*Männ***er***-chor*[17]
A + **e** + B	*Pferd***e***-stärke*[18]
A + **(-e+s)** + B	*Hilf***s***-morphem (Hilfe)*[19]
A + **(-e)** + B	*Schul-kind (Schule)*[20]
A + **ens** + B	*Mensch***ens***-kind*[21]
A + **(-en)** + B	*Schreib-maschine*[22]

Table 1: Connecting morphemes (Žepić, 1970).

Our compound generation module relies on the

[4] *state-contract, student-house*

[5] *health-office, cock’s feather*

[6] *diary, daylight*

[7] *nursery school, child’s head, child’s legs*

[8] A noun is categorized in an inflectional class based on its *genitive singular case* and *nominative plural number*. Often V is the same as an inflectional suffix of A.

[9] The zero morpheme, i.e. when the constituents are glued together without a specific connecting element, is represented with Ø.

[10] *insurance*

[11] *insurance contract*

[12] If, for example, A has a masculine gender and belongs to *inflectional class VI* (weak masc. nouns: gen. sg = *en*, nom. pl. = *en* as in (Žepić, 1970)), then it can take only one of the three morphemes *0*, *-(e)n*, or *-ens*, but not any of the other six morphemes. i.e. *-er*, *-(e)s*, etc. Nouns which belong to this class are *Mensch*, *Affe*, etc.

[13] A difference between paradigmic and unparadigmic connecting elements has been made in (Ortner et al., 1991; Fuhrhop, 1996; Fuhrhop, 1998). Because of this distinction, a finer classification is considered: for example, *-s* can appear both as paradigmic and unparadigmic connecting element, and therefore *-s* and *-es*, which are put together as a second morpheme in Table 1, may have different distributions.

[14] *steam engine*

[15] *season, inauguration speech*

[16] *electron tube*

[17] *men choir*

[18] *horsepower*

[19] *help morpheme*

[20] *pupil*

[21] *golly*

[22] *typewriter*

above mentioned linguistic analyses and empirical studies. Knowing all this information, the module tries to find the correct connecting element in the noun compound formation process. Section 3 describes our approach and implementation.

3 Concatenation of Compound Constituents

This section illustrates how we incorporated some of the linguistic knowledge presented in (Žepić, 1970; Ortner et al., 1991; Fuhrhop, 1996; Fuhrhop, 1998) in our machine translation system.

3.1 Data Preparation

As development and test data we used a subset of the GermaNet list of split nominal compounds (Henrich and Hinrichs, 2011), version 10.0, and compounds from Schulte Im Walde (2013). Table 2 shows how the GermaNet data of split noun compounds looks like. The first column contains the compounds. The second and third columns contain the base forms of the first and second compound constituents.

Compound	**C1**	**C2**
Ferienzeit [23]	Ferien	Zeit
Geschicht**s**buch [24]	Geschicht**e**	Buch
Dat**en**bank [25]	Dat**um**	Bank
Erdrotation [26]	Erd**e**	Rotation
...	...	...

Table 2: Split noun compounds (GermaNet).

Some of the first compound constituents have different forms in the compounds, e.g. *Datum* is transformed to *Daten*, *Geschichte* to *Geschichts*, and *Erde* to *Erd*. In order to prepare the *input (or source)* language for the compound generation module, we used the data from the second and third columns and simply concatenated the nouns in each line. Table 3 illustrates this process. The third column in Table 3 is our input (source) data.

Some of these input strings differ from the compounds in the first column of Table 2. The difference is in the form of the first compound constituent and the presence of a connecting morpheme inside the compound. With our compound generation module we *"translate (or map)"* from the concatenated base forms (noun constituents) to the correct compound forms. Table 4 represents the *source* and *target* strings.

C1	**C2**	**C1 concat. C2**
Ferien	Zeit	FerienZeit
Geschicht**e**	Buch	Geschicht**e**Buch
Dat**um**	Bank	Dat**um**Bank
Erd**e**	Rotation	Erd**e**Rotation
...	...	...

Table 3: Generating input data (third column).

Source		**Target**
FerienZeit	→	Ferienzeit
Geschicht**e**Buch	→	Geschicht**s**buch
Dat**um**Bank	→	Dat**en**bank
Erd**e**Rotation	→	Erdrotation
...		

Table 4: Source data (Table 3, column 3) and target data (Table 2, column 1).

3.2 Machine Translation

In order to generate the correct form of the compound constituents (Table 4, column 2) we used the Its-2 rule-based machine translation system (Wehrli et al., 2009).

3.2.1 Its-2

The compound generation module is a part of our Its-2 machine translation system. This system has a standard transfer architecture with parsing, transfer, and generation modules. The source language document is first analyzed by the Fips parser (Wehrli, 2007; Wehrli and Nerima, 2015), which segments the document into sentences, assigns to each sentence a phrase-structure representation and identifies the lexical units (words, idioms, collocations, etc.). The grammar of the Fips parser is free and personal adaptation of Chomskyan Generative Grammar (Chomsky, 1981; Chomsky, 1995). The transfer module maps the source-language representation into an equivalent target-language representation through a recursive traversal of the source-language structure in the order: head, left subconstituents, right subconstituents. Lexical transfer occurs at the head level. Language-pair specific transfer rules can alter the target language structure. The generation module is responsible for morphological and orthographical processes. The derivation of compounds takes place in the generation module.

[23] *holiday time* | *holiday* | *time*

[24] *history book* | *history* | *book*

[25] *database* | *date* | *base*

[26] *rotation of the earth* | *earth* | *rotation*

3.2.2 Lexical Database

The lexical database for Fips and Its-2 (Wehrli and Nerima, 2015) consists mainly of four types of lexicon: (*i*) the lexicon of **lexemes** (L), (*ii*) the lexicon of **words** (W), (*iii*) the lexicon of **collocations** (C), and (*iv*) the **bilingual lexicon** (BL). Table 5 presents the numbers of entries of the lexical database for English and German.

	L	W	C	BL
EN	58 332	105 025	9926	77 607
DE	43 854	451 140	3464	

Table 5: Lexical database for English and German.

Lexicalized Compounds In spite of the fact that there are good reasons to have a productive treatment of German compounds, there are some cases where the compounds must be lexicalized: when the generation of a compound is not predictable from the inflectional information of its constituents, all the spelling forms must appear in the lexicon of words and the compound must be represented as a lexeme. Another reason to lexicalize a compound is when its translation is non-compositional.

3.2.3 Compound Generation

As described in section 3.2.1, when Its-2 translates from one language to another, e.g., English to German, the source language is first analysed by the syntactic parser. For example, if we want to translate the string *history lesson*, the parser will produce the syntactic tree on Figure 1.

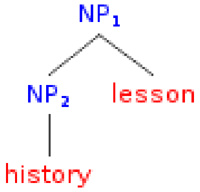


Figure 1: Syntactic analysis: source (EN).

During the transfer phase Its-2 maps the source syntactic structure into a corresponding target one, i.e. the structure on Figure 1 will be mapped to the syntactic structure on Figure 2.

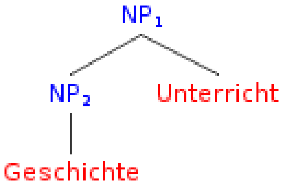


Figure 2: Transfer phase output: target (DE).

After the transfer phase we connect the two noun phrase branches of the syntactic tree produced after the transfer phase. In the current example, we obtain *GeschichteUnterricht*. This, however, is an incorrect German compound, because the first constituent *Geschichte* is in its base form, and it should be modified to *Geschichts*. The goal of our German generation module is to map the concatenated base forms to the correctly formed compound:

GeschichteUnterricht → *Geschichtsunterricht*

This process depends on the Its-2 lexical database and the compounding generation rules. If a compound is present in the database, it is recognized and output, and the generation process ends. If the compound is not in the database, the compounding generation rules are applied. The rules implement linguistic knowledge described in section 2.

Table 6 shows the 16 implemented compounding rules as used in the experiments presented in section 4. The names of the inflectional classes (e.g., *Infl. Class IV*, etc.) are as given in (Žepić, 1970). Some of the rules use limited lexical information (taken from (Ortner et al., 1991; Fuhrhop, 1996; Fuhrhop, 1998)), e.g. rule #2 uses a list of 9 simplex words (e.g., *Sicht, Nacht*)[27], rule #7 uses a list of 22 one-syllable nouns (e.g., *Amt, Sport*)[28], rule #8 uses a list of 13 nouns, which express some kind of *relation* with the other constituent (e.g., *Mutter, Bruder*)[29].

4 Evaluation

We implemented the compound generation module by using as development data about 900 noun compounds from GermaNet (Henrich and Hinrichs, 2011) and 158 noun compounds from the database of Schulte Im Walde (2013). We arbitrarily selected 344 other noun compounds from the GermaNet database as a test set. The Its-2 system had a problem to analyse/recognize 49 of

[27] sight, night
[28] agency, sport
[29] mother, brother

#1	IF B matches ($geber$ or $nehmer$), then $V = \emptyset$
#2	IF A ends in $simplex\_words$, and $length(A) > length(simplex\_words[i])$, then $V = s$
#3	IF A ends in (ung or $heit$ or etc.), then $V = s$
#4	IF A is feminine and A belongs to Infl. Class IV and A ends in e, then $V = n$
#5	IF A is neuter and A ends in en, then $V = s$
#6	IF $length(A) > 5$ and A is feminine and A ends in (t or d), then $V = s$
#7	IF A is *a known one-syllable noun*, then $V = s$
#8	IF A is masculine and A ends in er, and B is rel_noun, then $V = s$

#9	IF A is masculine and A belongs to Infl. Class VI { {IF A = $mensch$ and B = $kind$,
	then $V = ens$}, {IF A ends in e, then $V = n$}, {ELSE $V = en$ }}

#10	IF A is neuter and A ends in ($ment$ or at), then $V = en$
#11	IF A is feminine and A belongs to Infl. Class V and A ends in er, then $V = n$
#12	IF A has plural form with en and A ends in a, then $V = en$
#13	IF A ends in (um or us), then $V = en$ after subtracting the suffix
#14	IF A ends in $ling$ and A is masculine, then $V = s$
#15	IF A ends in ut and A is feminine, then $V = s$
#16	IF $length(A) < 7$ and A ends in el and A is masculine, then $V = s$

Table 6: Compounding rules.

these compounds[30], and this left us with a final test set of 295 noun compound[31].

4.1 German-to-German Translation

The evaluation has been set-up as a German-to-German translation task in Its-2. As input we had the concatenation of two compound constituents in their base forms (see Figure 2). The whole translation process is the one described in section 3.2.1, i.e. the input goes through parsing, transfer and generation. However, instead of two different languages, we consider German as both source and target language, with source language being the concatenation of two compound constituents in their base forms, and target language being the well-formed generated noun compound (see the example in section 3.2.3). That is, we map (potentially incorrect) concatenated base forms to well-formed noun compound.

4.2 Evaluation Results

Table 7 shows the evaluation results. As a baseline we considered the concatenation of two base forms without applying any modification. This is basically our source language (see Table 3). The results show that the compound generation module outperformed the baseline system by 16%.

	Accuracy
Baseline	61%
Its-2	77%

Table 7: Evaluation results.

	Correct	**Incorrect**	**Total**
Rules	204	68	272
DB	23	0	23
Total	227	68	295

Table 8: Numbers of correctly/incorrectly generated compounds by the Its-2 compounding rules and database.

In Table 8 we can see that from all 227 correctly generated compounds, only 23 (10%) were found in the lexical database and 204 (90%) were generated using the compounding rules. The compounding rules generated 204 (75%) compounds correctly, and 68 (25%) incorrectly. The error analysis showed that most of the errors come from (i) compounds, which have many similar features (e.g., same gender, inflectional paradigm, final sound), but use different connecting elements, for example, *Blume* → *Blumentopf*, but *Grenze* → *Grenzfläche* ; (ii) compounds which allow for more than one connecting element in similar context, e.g. *Jahresbilanz*, but *Jahrmillionen*. Such examples cause some of our rules to overgenerate.

There are several connecting elements, which are not covered by the compounding rules. This affects the evaluation results as well. The problem is that those elements are not deterministic (or al-

[30] For example, some constituents were unknown words to the system, and this did not provide a proper input set to the compound generation module. Our goal in this work was to evaluate only the performance of the compound generation module, i.e. whether it concatenates the constituents with the correct connecting morpheme. How well we translate a whole set of noun compounds from one language to another is another task and we will address it in future work.

[31] The second constituents in 7 of these compounds had slightly different inflectional forms from the original forms, e.g., *Fenster (window)* → *Fenstern (window (pl. dative))*, however, these changes did not affect the concatenation of the constituents and we left these 7 compounds in the test set.

most deterministic), and therefore we did not implement them. The most productive rules listed in Table 6 are those involving *-en*, *-n* and *-s*. There is also one special case for *-ens*. In general, the productive connecting morphemes in German are *-e*, *-er*, *-en*, *-n* and *-s* (Fuhrhop, 1998). *-e* is used often with different first noun constituents in different ways, and it is therefore difficult to predict. The use of *-er* depends very much on semantics, i.e. whether the meaning of the first constituent is singular or plural.

5 Future Work

The statistical approaches to noun compound translation are prevalent (Rackow et al., 1992; Popovic et al., 2006; Stymne, 2008; Stymne et al., 2013). There is not much work done on generation of noun compounds into Germanic languages (Stymne and Cancedda, 2011). German noun compounds, which have first constituents allowing for multiple connecting elements, are still a big challenge. In the rule-based machine translation framework two possible solutions to this problem can still be explored: (*i*) enriching the lexical database of the machine translation system; (*ii*) discovering more rules and restrictions, which can give us the knowledge of how to use the connecting elements properly. The latter is quite ambitious and would require a joint effort of researchers from different areas (theoretical linguists, computational linguist, etc.).

Acknowledgements

This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS). We would like also to thank to Sabine Schulte im Walde from the University of Stuttgart, for providing us with their noun compound data.

References

N. Chomsky. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

N. Chomsky. 1995. *The Minimalist Program*. Cambridge, MA: The MIT Press.

P. N. Garner, R. Clark, J.-P. Goldman, P.-E. Honnet, M. Ivanova, A. Lazaridis, H. Liang, B. Pfister, M. S. Ribeiro, E. Wehrli and J. Yamagishi. 2014. *Translation and Prosody in Swiss Languages*. Nouveaux cahiers de linguistique française (31), 3rd Swiss Workshop on Prosody, Geneva.

N. Fuhrhop. 1996. Fugenelemente. *In: E. Lang and G. Zifonun eds. 1996. Deutsch typologisch (IDS Jahrbuch 1995)*. Berlin/New York: de Gruyter. pp.525-550.

N. Fuhrhop. 1998. *Grenzfälle morphologischer Einheiten*. Tübingen: Stauffenburg Verlag.

V. Henrich and E. Hinrichs. 2011. Determining Immediate Constituents of Compounds in GermaNet. *In Proceedings of RANLP*. pp. 420-426.

L. Ortner, E. Müller-Bollhagen, H. Ortner, H. Wellmann, M. Pümpel-Mader, and H. Gärtner. 1991. Deutsche Wortbildung, Haupttl.4, Substantivkomposita: Typen Und Tendenzen In Der Gegenwartssprache: 4. Haupttl (Komposita Und Kompositionsahnliche Strukturen 1). *Gruyter*.

C. P. Escartín, S. Peitz and H. Ney. 2014. German Compounds and Statistical Machine Translation. Can they get along? *In Proceedings of the 10th Workshop on Multiword Expressions*.

M. Popovic, D. Stein and H. Ney. 2006. Statistical Machine Translation of German Compound Words. *LNCS, Springer*.

U. Rackow, I. Dagan and U. Schwall. 1992. Automatic Translation of Noun Compounds. *In Proceedings of COLING*.

Rat für deutsche Rechtschreibung. 2006. Deutsche Rechtschreibung: Regeln und Wörterverzeichnis: Amtliche Regelung. *Tübingen: Gunter Narr Verlag*.

S. Stymne. 2008. German Compounds in Factored Statistical Machine Translation. *In Proceedings of the 6th International Conference on Natural Language Processing (GoTAL-08)*.

S. Stymne and N. Cancedda. 2011. Productive Generation of Compound Words in Statistical Machine Translation. *In Proceedings of the Sixth Workshop on SMT* Edinburgh, Scotland. 250–260.

S. Stymne, N. Cancedda and L. Ahrenberg. 2013. Generation of Compound Words in Statistical Machine Translation into Compounding Languages. *Computational Linguistics 39(4)* 1067-1108.

S. Schulte Im Walde, S. Müller and S. Roller 2013. Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. *In Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM). Atlanta, GA.*

E. Wehrli. 2007. Fips, a "Deep" Linguistic Multilingual Parser. *In Proceedings of the Workshop on Deep Linguistic Processing, Prague, Czech Republic*. pp. 120–127.

E. Wehrli, L. Nerima and Y. Scherrer. 2009. Deep Linguistic Multilingual Translation and Bilingual Dictionaries. *In Proceedings of the Fourth Workshop on SMT*.

E. Wehrli and L. Nerima. 2015. The Fips Multilingual Parser. *In Language Production, Gognition and the Lexicon, Gala N., Rapp R. and Bel-Enguix G. Eds., Springer*, pp. 473–490.

S. Žepić. 1970. *Morphologie und Semantik der deutschen Nominalkomposita*. Izdavački zavod Jugoslavenske akademije znanosti i umjetnosti, Zagreb.

Multi-Word Expressions in User-Generated Content: How Many and How Well Translated? Evidence from a Post-editing Experiment

Violeta Seretan
Department of Translation Technology
Faculty of Translation and Interpreting, University of Geneva
40 Bvd. du Pont-d'Arve, CH-1211 Geneva, Switzerland
Violeta.Seretan@unige.ch

Abstract

According to theoretical claims, multi-word expressions are pervasive in all genres and domains, and, because of their idiosyncratic nature, they are particularly prone to automatic translation errors. We tested these claims empirically in the user-generated content domain and found that, while multi-word expressions are indeed common in this domain, their automatic translation is actually often correct, and only a modest amount – about one fifth – of the post-editing effort is devoted to fixing their translation. We also found that the upperbound for the increase in translation quality expected from perfectly handling multi-word expressions is 9 BLEU points, much higher than what is currently achieved. These results suggest that the translation of multi-word expressions is nowadays largely correct, but there is still a way to go towards their perfect translation.

1 Introduction

The literature on multi-word expressions (henceforth, MWEs) abounds with claims on the pervasiveness of such expressions in language, as well as on the difficulty of translating these expressions automatically. It is held that multi-lexeme units are of the same number of magnitude as single-lexeme units (Jackendoff, 1997), or even one order of magnitude more numerous (Mel'čuk, 1998). Also, it has been suggested that no single utterance is totally free of MWEs (Lea and Runcie, 2002) and that MWEs are a major obstacle for achieving correct machine translation (Sag et al., 2002).

In order to cope with the MWE translation problem, the solutions adopted in the literature were to gather MWEs in the lexica of rule-based translation systems, as in Orliac and Dillinger (2003), or to find means to integrate them into the statistical machine translation (SMT) pipeline. In the latter case, the MWE integration is achieved either in the pre-editing stage in a so-called *words-with-spaces* approach (Carpuat and Diab, 2010), or in the training stage as supplementary "sentences" (Bouamor et al., 2012), or, again, by adding MWEs to the phrase table together with new features that constrain the SMT decoder to apply a MWE-compatible segmentation over a different one (Carpuat and Diab, 2010).

These integration strategies were reported to yield positive results. For instance, Orliac and Dillinger report "significant improvement in readability and perceived quality of the translation produced" (Orliac and Dillinger, 2003, 293). Also, a number of authors (see Table 1) report significant improvements of SMT quality in terms of BLEU score. The increase in translation quality which is due to MWE integration remains, however, quite limited with respect to the whole sentence score. As can be seen in Table 1, the increase is often less than one BLEU point. This modest increase seem to contradict the original statements on the pervasiveness of MWEs and their importance for achieving better translations.

The aim of our study is to look more in detail at the issue of MWE translation, in order to better understand the reasons beyond this positive but limited impact observed, and beyond the apparent contradiction with theoretical claims. First, we wanted to check whether MWEs are as pervasive in our data domain (the user-generated content domain) as the literature claims. Second, we wanted to investigate how many of the MWEs are badly translated. Third, we wanted to to see how much we could gain in terms of translation quality score if we had a perfect, 'oracle' translation for all MWEs in our test set (i.e., to determine the upperbound that a system could achieve, compared to the state of the art of one BLEU point).

Work	Language Pair	Test Set	Impact (BLEU Points)
Bai et al. (2009)	cn–en	NIST MT-06	0.22 (21.79 – 21.57)
Carpuat and Diab (2010)	en–ar	NIST MT-08	0.78 (31.27 – 30.49)
Liu et al. (2010)	cn–en	NIST MT-04	0.85 (30.47 – 29.62)
Tsvetkov and Wintner (2010)	he–en	not reported	0.10 (13.79 – 13.69)
Liu et al. (2011)	cn–en	NIST MT-04	1.10 (29.87 – 28.77)
	cn–en	NIST MT-2008	1.41 (19.83 – 18.42)
Bouamor et al. (2012)	fr–en	Europarl	0.30 (25.94 – 25.64)
Kordoni and Simova (2014)	en–bg	SeTimes	0.20 (23.9 – 23.7)

Table 1: Impact of multi-word expression integration on translation quality.

Our hypotheses, which were grounded on theoretical research, were the following:

1. MWEs are pervasive in user-generated content;
2. Most MWEs are badly translated;
3. Because of the above, post-editors spend a lot of effort correcting MWE translation errors.

We conducted empirical investigation on post-editing data available from the ACCEPT European project devoted to improving the translatability of user-generated content.[1] This data domain is less explored by existing MWE research, despite the fact that it represents one of the biggest challenges for natural language processing for the years to come. Our study provides evidence for the prevalence of MWEs in the social media genre represented by forum posts. While the pervasiveness assumption was confirmed, the other assumptions were challenged.

In the following sections, we describe the data (Section 2) and the investigation these data allowed, referring to above-mentioned hypotheses and to the findings obtained (Section 3). In the last section, we provide concluding remarks and ideas for future work (Section 4).

2 Data

The data used in our study is taken from a larger dataset available from the ACCEPT European projet (2012–2014) devoted to improving SMT of user-generated content. The dataset consists of 1000 technical forum posts in French, which have been automatically translated into English using a domain-adapted phrase-based statistical machine translation system (D41, 2013). The MT output has been manually corrected by a post-editor, a native speaker of English, paid for the task. The forum posts originate from the French chapter of the Norton Community Forum related to computer security issues.[2]

From the total 4666 corresponding translation segments, we randomly sampled 500 segments for the purpose of this study. Three of these segments turned out to be in English, as they were quotes of error messages included by forum users in their posts. After discarding these segments, we ended up with 497 segments in our test set. They total 5025 words and their length varies between one word (e.g., *Bonjour*, 'Hello') and 73 words, with an average size of 10.1 words/segment.

The example below shows a segment, its automatic translation, and the version corrected by the post-editor:

(1) a. *Source: Laissez tomber* ..... depuis 5 mois ..... j 'ai *résolu* la *question* hier
b. *MT: Let down* ..... for 5 months ..... I've *resolved* the *issue* yesterday
c. *Post-edited: Drop it* ..... after 5 months ..... I *fixed* the *issue* yesterday.

This example illustrates the discussion in Section 1. There are two MWEs in this segment, shown in italics. Both are badly translated and, as can be seen, their correction represent a large share of the total amount of corrections made by the post-editor.

3 Experiments and Results

In order to test the validity of the hypotheses put forward in Section 1, we conducted a series of experiments, summarised below.

[1] www.accept-project.eu. Accessed July, 2015.

[2] http://fr.community.norton.com. Accessed July, 2015.

3.1 Checking MWE pervasiveness

The assumption that MWE are pervasive in language is often taken for granted in the literature, as it seems superfluous to demonstrate the obvious. There are, however, studies in which authors provide empirical evidence supporting this claim. For instance, Howarth and Nesi (1996) came to the conclusion that "most sentences contain at least one collocation" (Pearce, 2001). While this holds for the general domain, little is known about the validity of this claim for the user-generated content domain.

In order to check the pervasiveness of MWEs in this domain, we proceeded to the manual identification of all MWEs in our test set of 497 segments. There are tools for MWE identification in text, or dictionaries we could have used to recognise MWEs in text. But we have chosen to annotate MWEs manually, mainly for the reason that user-generated content exhibit peculiarities which hinder the application of automatic methods: presence of slang, abbreviations, colloquial speech, errors at various levels (punctuation, casing, spelling, syntax, style, etc). Users of technical forums like the Norton Community forum are most likely to write in a hurry, because they are concerned by their problem at hand – for instance, by the fact that their computer crashes all the time, or the fact that some product they installed keeps on debiting their card monthly despite the subscription being cancelled. Their real concern is getting a solution to their problem as soon as possible, not the quality of their message. Any deviation from the norm is acceptable, as long as the message is understood by the community members.

Therefore, we chose to do the MWE annotation entirely manually, this way ensuring the accuracy of the annotation. The criterion used in deciding weather a combination is a MWE is the lexicographic criterion, i.e., we annotated a combination as MWE *iff* it was deemed worth of inclusion in a lexicon (in other words, it was not a regular combination). Despite this simple criterion, there is always some amount of uncertainty and subjectivity, as it is a well-known fact that MWEs are on a continuum from completely regular to completely idiosyncratic, and it is impossible to draw a clear-cut line between regular combinations and combinations which are MWEs (McKeown and Radev, 2000). In future studies, we may want to rely on judgements from multiple annotators in order to reduce the amount of uncertainty and subjectivity.

A specific annotation choice was necessary in the case of nested MWEs, i.e., when a MWE participates in another MWE. An exemple is provided below, in which *mise à jour* (lit., put to day, 'update') further combines with the verb *faire* to form a longer MWE, *faire mise à jour* (lit., to do update, 'to update'):

(2) Malgré les mises à jour faites (Démarrer>Windows Update), windows demande toujours les mêmes 2 maj

In this case, the decision taken was to count each MWE instance separately. Therefore, in this examples, we counted two MWEs.

Another specific annotation choice concerned the annotation of MWE reduced to abbreviations. In Example (2) above, there is a second instance of the MWE *mise à jour* occurring at the end of the sentence, as the abbreviation *maj*. In the framework of the ACCEPT project from which the data derives, abbreviations were treated as non-standard lexical items that have to be normalised in order to facilitate translation. As can be seen in Example (3), the post-editor understood the French abbreviation and corrected the MT output by proposing the full form equivalent in English, *update*. Influenced by the pre-editing approach adopted in the context of the project (*maj* → *mise à jour*), we decide to count abbreviations of MWEs as actual MWE instances.[3]

(3) a. *MT:* Despite updates made (Start > Windows Update), windows always ask the same 2 Shift

b. *Post-editor:* Despite the updates done (Start > Windows Update), windows always asks for the same two updates.

Given the methodological choices explained above, the statistics for the test set are as follows. The total number of MWEs in the 497 segments is 223. A number of 152 segments contain MWEs, which gives an average of 1.5 MWEs/segment. This might seem in line with known results from literature; however, reported to the total test size, the average is 0.4, lower than stipulated by litera-

[3] As an alternative, we could have ignored abbreviations, as one workshop attendee suggested. We maintain, however, that in a translation perspective, contracted MWEs require a full form version in order to facilitate their treatment.

ture. Since many segments are very short, we ignored segments that contained less than 100 characters, and got an average of 1.3 MWEs/segments for the remaining 91 segments.

Our results indicate that in the user-generated content domain, there seem to be less MWEs than in the general domain. However, the words participating in MWEs make up as much as 10.5% of the total words in the test set. Previous results for the general domain reported that MWEs account for just only 5% of the data in the NIST-MT06 test set (Bai et al., 2009). While a straight comparison is not possible because of the different methodologies used to recognise MWEs, the relatively high percentage obtained for the user-generated content domain suggests that MWE account for a larger portion of the data. From a translation perspective, it is important to focus on this portion of the data because it is likely to be more important in terms of comprehensibility of the MT output.

3.2 Checking MWE Translation Quality

The question arise if the automatic translation of MWEs requires any correction, in the first place. As Babych observes, “SMT output is often surprisingly good with respect to short distance collocations” (Babych et al., 2012, 103). Good translations for idiomatic expressions can still be achieved in SMT as a by-product of learning from parallel corpora. This can be seen, for instance, in Example (4)). The MT output required no correction at all from the post-editor.

(4) a. Faites-nous part de vos expériences
b. Please email us your experiences

Example (5), on the contrary, shows a bad translation. The collocation *rencontrer erreur* is translated literally by the system.

(5) a. *Source:* Je viens de *rencontrer* une *erreur* à l ’instant en faisant un Live Update manuel
b. *MT:* I have just *met* an *error* just now by a Live Update manual
c. *Post-editor:* I have just *had* an *error* just now doing a manual Live Update

To report the number of MWEs that are well translated by the system, we relied on the post-editor’s version as a gold standard. Whenever the editor changed the MWE translation as proposed by the system, we considered it was wrong, except for the cases where the changes were minor, like fixing number or agreement.

According to this method, the percentage of well-translated MWEs is 63.2% (141/223). Therefore, less than half of MWEs required a different translation. This result contradicts our expectations induced by theoretical claims, which would predict a higher rate of failure. It might also explain the limited impact of MWE integration observed in the literature: if MWEs account for about 5% of the data and more than half are well translated anyway, the small increase in BLEU seems justified.

Previous research (Bod, 2007; Wehrli et al., 2009; Babych et al., 2012) has suggested that SMT is more problematic for the more flexible expressions. This problem is exacerbated in our domain, as shown in Example (6). The SMT system fails to correctly translate the MWE *mise à jour* because its form deviates from the expected form and takes an unconventional plural form:

(6) a. *Source: mise à jours* live update
b. *MT: Upgrade days* live update
c. *Post-editor: Update* to live Update .

Due to time constraints, for the present experiment we did not relate yet the quality of MWE translation to the flexibility of the expressions, in order to find wether there is an effect. This analysis is left for future work. We tested, however, the statistical significance of the difference between the total number of MWE in the 152 segments containg MWEs, on the one hand, and the number of correctly translated MWEs, on the other hand. This difference is extremely significant ($t(151) = 9.93, p < 0.001$). This means that the problem of MWEs in translation is real. A significant number of MWEs are badly translated. If we focus on MWEs, we only deal with about 10% of the data (see Section 3.1), but arguably we deal with the most critical portion of the data compared to other corrections which might not be as critical. Fixing a determiner, number or agreement might not have the same impact on comprehensibility as fixing a collocate (see Example (5)). Moreover, MWE translation errors seem to make a large share of all errors, because MWEs are common and they are often badly translated. This hypothesis is tested in the experiment described next.

	BLEU	WER	TER	Levenshtein
Total effort (MT)	0.511	0.316	0.291	24.0
Effort excluding MWE correction (Oracle)	0.603	0.249	0.225	19.8
Effort spent on MWEs (difference)	0.092	-0.066	-0.066	4.2
Effort spent on MWEs (%)	18.0%	-21.0%	-22.6%	17.6%
t(151)	-6.83	12.25	6.16	8.46

Table 2: Post-editing effort spent fixing MWE translation errors.

3.3 Quantifying MWE Correction Effort

How much of the total post-editing effort is actually spent on fixing MWE translation errors? To answer this question, we quantified the post-editing effort in terms of standard metrics used in the field (BLEU, TER, WER, Levenshtein).[4] We compared the total post-editing effort against the post-editing effort excluding MWE correction. The difference represents the effort devoted to fixing MWE translation errors.

The total post-editing effort is, obviously, computed for the MT output as such. The effort excluding MWE correction is computed on a modified version, on which the correct, 'oracle' MWE translation is extracted from the gold standard, which is the post-editor's version. To illustrate this, we provide an example below (Example (7)). The MWE correction *tried all ways* is inserted from the post-editor's version, while the rest is left unchanged (notice the post-editor further changed *do not* into *can't* at the end of the sentence).

(7) a. *Source:* J'ai *retourné* le programme *dans tout les sens* pour trouver l'option qui permet de changer le mot de passe mais je ne la trouve pas.

b. *MT:* I have *returned* the program *in any sense* to find the option that lets you change the password but I do not find it.

c. *Post-editor:* I have *tried all ways* to find the option that lets you change the password but I can't find it.

d. *Oracle:* I have *tried all ways* to find the option that lets you change the password but I do not find it.

The results are shown in Table 2. MWEs account for about a fifth of the total post-editing effort, according to the metrics used. Admittedly, this is less than expected considering theoretical arguments. Like the bad translation hypothesis, the hypothesis that most of the post-editing effort is focused on MWEs is invalidated by our study. However, this result is based on the selection of metrics used to quantify the post-editing effort. We believe that there are more accurate metrics of measuring effort, like time. Had we had time logs for our data, we could have come up with a different conclusion. Indeed, the time needed for providing a correct translation for a MWE is arguably much longer than the time required to delete a determiner of to fix agreement issues. Further investigation is therefore needed in order to reliably invalidate the hypothesis in question.

As for the statistical significance of results, the difference between the total effort and the effort excluding MWEs correction is extremely significant ($p < 0.001$), as can be seen in the last row of Table 2. This means that the MWE correction effort is significant. Again, the interpretation of this finding is that MWEs constitute a real problem for machine translation.

It is important to note that if MWEs are handled perfectly, the expected increase in translation quality can be as high as 9.2 BLEU points, while current integration methods achieve about 1 BLEU point, as seen in Section 1.

4 Conclusion

Summing up, while the literature put emphasis on the prevalence on the prevalence of MWEs and their importance for translation, little was known about the empirical validity of theoretical claims, and even less so about their validity in the specific domain of user-generated content. This domain is little investigated by MWE research, but is of major interest for natural language processing in general and for machine translation in particular.

[4] BLEU measures the distance from a reference translation at the word level, using n-grams. TER measures the same distance in terms of operations at the word level (substitution, insertion, deletion, shift). WER is similar to TER, with no shift. The Levenshtein distance on which TER and WER are based works similarly, but at the character level. Other post-editing effort measures are the time and keystrokes. Time and keystoke logs are unfortunately not available for our data.

The aim of the present study was to test the validity of theoretical claims for this domain, in order to find out, in particular, how frequent MWEs and MWE translation errors really are, and how much of the total post-editing effort is spent on correcting MWE translation errors. We conducted a study based on large-scale post-editing dataset, which allowed us to validate the MWE prevalence assumption and to find out that MWEs account for more than 10% of words in our dataset. We also checked the bad translation assumption and found that the majority of MWEs are actually correctly translated. This is different from what the literature suggests, but we found that the number of badly-translated MWEs is, however, significant. As for the integration of MWE knowledge into MT systems, we computed an upperbound for the increase in translation quality we could expect by better handling MWEs: if we handle them perfectly, we could gain as much as 9 BLEU points. These results suggest that there is still room for improvement in this area.

This study could be extended to more language pairs and new datasets, by exploiting multiple annotations, and quantifying the MWE translation correction effort in terms of time, in addition to automatic metrics.

References

Bogdan Babych, Kurt Eberle, Johanna Geiß, Mireia Ginestí-Rosell, Anthony Hartley, Reinhard Rapp, Serge Sharoff, and Martin Thomas. 2012. Design of a hybrid high quality machine translation system. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 101–112, Avignon, France, April.

Ming-Hong Bai, Jia-Ming You, Keh-Jiann Chen, and Jason S. Chang. 2009. Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 478–486, Singapore.

Rens Bod. 2007. Unsupervised syntax-based machine translation: the contribution of discontiguous phrases. In *Proceedings of MT Summit XI*, pages 51–56, Copenhagen, Denmark, September.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, California, June. Association for Computational Linguistics.

2013. ACCEPT deliverable D 4.1: Baseline MT systems. `http://www.accept.unige.ch/Products/D_4_1_Baseline_MT_systems.pdf`.

Peter Howarth and Hilary Nesi. 1996. The teaching of collocations in EAP. Technical report, University of Leeds, June.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Valia Kordoni and Iliana Simova. 2014. Multiword expressions in machine translation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Diana Lea and Moira Runcie, editors. 2002. *Oxford Collocations Dictionary for Students of English*. Oxford University Press, Oxford.

Zhanyi Liu, Haifeng Wang, Hua Wu, and Sheng Li. 2010. Improving statistical machine translation with monolingual collocation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 825–833, Uppsala, Sweden, July.

Zhanyi Liu, Haifeng Wang, Hua Wu, Ting Liu, and Sheng Li. 2011. Reordering with source language collocations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1036–1044, Portland, Oregon, USA, June.

Kathleen R. McKeown and Dragomir R. Radev. 2000. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *A Handbook of Natural Language Processing*, pages 507–523. Marcel Dekker, New York, USA.

Igor Mel'čuk. 1998. Collocations and lexical functions. In Anthony P. Cowie, editor, *Phraseology. Theory, Analysis, and Applications*, pages 23–53. Claredon Press, Oxford.

Brigitte Orliac and Mike Dillinger. 2003. Collocation extraction for machine translation. In *Proceedings of Machine Translation Summit IX*, pages 292–298, New Orleans, Lousiana, USA.

Darren Pearce. 2001. Synonymy in collocation extraction. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 41–46, Pittsburgh, USA.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Coling 2010: Posters*, pages 1256–1264, Beijing, China, August.

Eric Wehrli, Violeta Seretan, Luka Nerima, and Lorenza Russo. 2009. Collocations in a rule-based MT system: A case study evaluation of their translation adequacy. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*, pages 128–135, Barcelona, Spain.

Transformation and Multiword Units in Quechua

MAXIMILIANO DURAN
Université Franche Comte
Besançon France
duran_maximiliano@yahoo.fr

Abstract

This article presents the process of how with the aid of the transformational engine of the NooJ[1], linguistic development environment, we may identify, annotate and transform Quechua[2] MWU; generate paraphrases for a given Lexicon-Grammar class of MWU sentences, taking into account the grammatical restrictions of the applicability of such transformations.

1 Introduction

An important number of the multi word units, MWU, in Quechua are collocations of two-PoS[3] units. There are at least 12 classes of them: N_N, N_V, V_N, A_A…. The resulting MWU may be a new noun, a new adjective or a new verb as we can see in the following table:

N-N>A *qari qari*> without fear (*qari* : man)
N-N>N sacha sacha> a forest (*sacha*: tree)
N-V>ADV *wallpa waqayta*> at twilight (*wallpa*: hen, *waqay(ta)*: to cry)
N-ADV>A *runa masinchik* > our fellow human (*runa*: human, *masi(nchik)*: similar)
N_V(na)[4]> A *anku chutana*>steep path (*anku*: Achilles tendon, *chuta(na)*: to stretch*)*
V_N>N *samai wasi*>guest house (*samay*: to rest, *wasi*: house)
A-A> A *yuraq yuraq*> very white (*yuraq*: white)

A-N>A *raku kunka*> baritone (*raku*: thick, *kunka*: neck)
ADV-V>ADV *hina kachun* >o.k. (*hina*: similar, *ka(chun)*: to be)
A-N>ADV *huk similla*>unanimously (*huk*: one, *simi(lla)*: mouth*)*
ADV-ADV> ADV *qawanpi ukunpi*>chaotically (*qawa(npi)*: outside, *uku(npi)*: inside*)*
V_V>N *mikuchikui upyachikui* wedding party (*miku(chikui)*: to eat, *upya(chikui)*: to drink

According to the Quechua grammar, any noun and any adjective may be duplicated. In general the duplication of adjective yields as a result the superlative of the adjective and the duplication of a noun may mean an important increase in number of the noun or a change of the semantic field of the resulting MWU.

I remark that when we symbolize N, V or A, I actually make reference to the paradigm symbolized by N which include the noun but also a certain class of its inflected forms (as we show some lines later). The same remark is applied to the other symbols.

There exists also many MWU made of distinct noun components N1, N2 acting with a particular type of verbs.

N1_N2(n-wan) *maymanpas ustuchkan*
qara uya*-n-wan maymanpas ustu-chka-n* he is sticking his nose anywhere as a rascal
qara uya*-n-wan maymanpas ri-chka-n* anywhere he goes he is sticking his nose as a rascal
qara uya*-n-wan maymanpas yayku-chka-n* he is entering anywhere to stick his nose as a rascal

The dictionary of verbs involved in this pattern is made of verbs of movement V_MO1={*ustuy, riy, yaykuy, paway, puriy,…*}.

N1_N2(ka-spa) maytapas V_MO1+CHKA +PR+s+3
qara uya *ka-spa maytapas(ustu-chka-n)* shameless as he is, he sticks his nose anywhere
qara uya *ka-spa maytapas rin (ri-chka-n)* shameless as he is, he goes anywhere to stick his nose
qara uya *ka-spa maytapas yayku (yayku-chka-n)* shameless as he is, he sticks his nose anywhere

A pattern including negation: **N1_N2** ka-spa mana V_CO1 ku+PR+s+3+chu

where the verbs concerned are a the behavioral class CO1 {*penqay, manchay,…*}

[1] Silberztein, M.(2003) NooJ Manual. htpp://www.nooj4nlp.net (220 pages updated regularly).
[2] The Quechua language was the official language of the Inca civilization in Peru
[3] POS part of speech
[4] V(*na*) nominalization with *na* of the verb V

And let us present some examples of MWU including other PoS like N_V, V_N or V1_V2 components:

In the case **N_V(na)>A**. If the resulting MWU is an adjective, like in example that follows, it can be inflected using the nominal paradigms applied to the last component of the MWU (this component is a verb but in its nominal form because it contains the suffix -*na*):

anku chutana>steep (a hill, a path) (*anku*: Achilles tendon(N,) *chuta(na)*: to stretch(V))

Pablo rin ***anku chutana*** *ñanninta* Pablo walks by the steep path

Pablo rin ***anku chutanan-ta*** Pablo walks by the steep one

anku chutanan-ta-chu *Pablo rin?* *Does* Pablo walks by the steep path?

Similarly, in the case **V(q)_N >A** (*pasaq simi*). The resulting adjective can also be inflected using the nominal/adjectival paradigms (the verb is in nominal form because it contains the suffix -*q*) applied to the last component of the MWU.

pasaq simi squealer (***pasa****(q)*: to pass(V,) ***simi***: mouth (N))

Pablo willaikun ***pasaq simi*** wauqinman Pablo has told it to his squealer brother

pasaq simi-ta Pablo niikun Pablo has told it to the squealer one

pasaq simi-ta-chu qwarqanki? Have you seen the squealer one?

The MWU resulting from the collocation of two verbs in which V1(i)_V2(i)>N can be transformed by means of the inflections of the second component.

miku(chikui) upya(chikui): wedding party **mikuy**: to eat

pablom rimachkan ***mikuchikui upyachikui****-nin-manta* Pablo talks about his wedding party

pablom ***mikuchikui upyachikui****-nin-manta rimachkan* Pablo is talking about his wedding party

pablom ***mikuchikui upyachikui****-nin-manta rimarqan* Pablo has talked about his wedding party

pablom ***mikuchikui upyachikui****-nin-manta rimanqa* Pablo will talk about his wedding party

Knowing that the morpho-syntactic behavior of the components of a MWU influences on its morphology, we need to take a glance on the inflections and derivations of a Noun, an Adjectif, an ADVerb and a Verb to better identify and manage the MWU.

2 The Quechua Noun Inflection

Because of the rich inflection system of Quechua, we can see that a single Quechua form replaces a whole English phrase

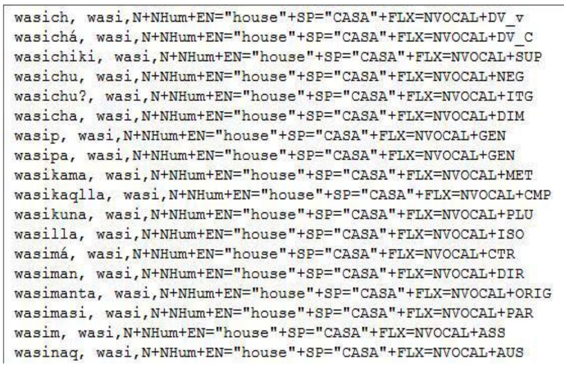
wawanchikraikullapas : Let us do it at least having in mind also our child

(nominal root: *wawa* child and *-nchik -raiku –lla and – pas* are nominal suffixes having their own semantic value[5] which explicit the global sense of the form).

The agglutination of suffixes. The 68 nominal suffixes[6] maybe attached to the noun as a single suffix or in combinations of two, three or more of them (generally up to 8 suffixes) in order to obtain an inflected form of the noun, like in the examples:

Noun	*wasi*	house
Suffixes :	*-nchik*	POS + p+1
	-pas	including, also
	-kuna	plural

Which gives us the following inflections: *wasi-kuna* the houses; *wasi-kuna-pas* including the houses; *wasi-nchik-kuna-pas* including our houses



```
wasich, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+DV_v
wasichá, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+DV_C
wasichiki, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+SUP
wasichu, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+NEG
wasichu?, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+ITG
wasicha, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+DIM
wasip, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+GEN
wasipa, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+GEN
wasikama, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+MET
wasikaqlla, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+CMP
wasikuna, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+PLU
wasilla, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+ISO
wasimá, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+CTR
wasiman, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+DIR
wasimanta, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+ORIG
wasimasi, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+PAR
wasim, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+ASS
wasinaq, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+AUS
```

Figure 1. A sample of the 3094 inflected forms of the noun *wasi.*

Using the corresponding paradigms programmed in NooJ we obtain 3094 inflected forms of wasi (using combinations of less than four suffixes) as shown in Figure 1.

[5] We have built as a NooJ linguistic resource, a dictionary of the semantic values of all the nominal, adjectival and verbal suffixes called Qu_Suff_sem.dic. See Annex 1)

[6] **Suff_N** *={-ch , chá, -cha , -chiki, -chu, -chu?, -p, -pa, -kama, -kaqlla, -kuna, -lla, -má, -man, -manta, -masi, -m, -mi, -mpa, -nimpa, -naq , -nta , -ninta , -nintin, -ntin, -niraq, -niyuq , -ña, -niq, -p , -pa, -paq , -pas, -pi, -poss(7v+7c), -puni, pura, -qa, -hina, -raiku, -raq , -ri , -sapa , -s, -si , -su, -ta, -taq, -wan, -y!, niy!, -ya!, -yá, -yupa , -yuq} (68)*

Where (7v,+7c) is the set seven possesive suffixes poss (7v) = (–i, -iki, -n, -nchik, -iku, -ikichik, -nku) for the vowel endings ; and the set poss (7c) = (–nii, -niiki, -nin, -ninchik, -niiku, -niikichik, -ninku) for the consonant endings.

An adjective, a pronoun or an adverb can be inflected by the same set of suffixes Suff_N. e.g.: *puka* red will become (A) *puka-pas*: also the red; *puka-man*: towards the red; (PRO) *ñuqa(paq*for me; (ADV) *hina(man)* (*hina*: similar) towards the similar one, etc. These inflected forms may be collocated to get MWU as we will see later.

3 The Quechua Verbal Forms

The lexicon of simple verbs is small (<1400) and yet the numerous nominalizations and derivations of verbs and composed verbs increase this number considerably as we have shown in a recent work (Duran, 2015) to appear[7].

On conjugations all the verbs are regular. For this we have programmed only one NooJ grammar for the paradigm of conjugation of the present tense for all the verbs as follows.

PR=*<B>(ni/PR+s+1|nki/PR+s+2|n/PR+s+3|nchik/PR+pin+1|nkichik/PR+p+2 |nku/PR+p+3 |niku/PR+pex+1);*

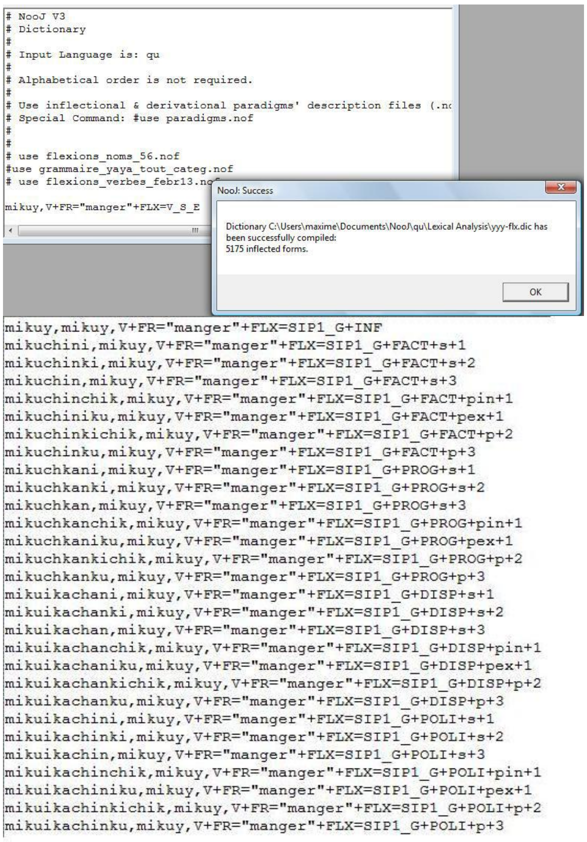
Moreover the present tense is a key structure for the realization of many verbal forms having a rapport with the present (past, past participle, gerund, etc)[8].

The personal ending in the conjugation of the present acts as a fixed point in the transformations as follows

*rima-**nchik*** we talk
*rima-ri-chka-**nchik*** we are beginning to talk
*rima-ri-chka-**nchik** –ña* we are already beginning to talk

Let us notice that certain suffixes appear before the ending ***nchik***, we call them Inter posed suffixes (IPS)[9] and others like *ña*, in this example, after the ending ***nchik*** (p+1), we call them post posed suffixes (PPS)[10].

Parsing the corresponding paradigms of inflections programmed in NooJ we obtain 5175 inflected forms of the verb *mikuy:* to eat (with combinations of less than four suffixes) as it appears in Figure 2.



```
# NooJ V3
# Dictionary
#
# Input Language is: qu
#
# Alphabetical order is not required.
#
# Use inflectional & derivational paradigms' description files (.nc
# Special Command: #use paradigms.nof
#
#
# use flexions_noms_56.nof
#use grammaire_yaya_tout_categ.nof
# use flexions_verbes_febr13.no

mikuy,V+FR="manger"+FLX=V_S_E
```

NooJ: Success

Dictionary C:\Users\maxime\Documents\NooJ\qu\Lexical Analysis\yyy-flx.dic has been successfully compiled:
5175 inflected forms.

OK

```
mikuy,mikuy,V+FR="manger"+FLX=SIP1_G+INF
mikuchini,mikuy,V+FR="manger"+FLX=SIP1_G+FACT+s+1
mikuchinki,mikuy,V+FR="manger"+FLX=SIP1_G+FACT+s+2
mikuchin,mikuy,V+FR="manger"+FLX=SIP1_G+FACT+s+3
mikuchinchik,mikuy,V+FR="manger"+FLX=SIP1_G+FACT+pin+1
mikuchiniku,mikuy,V+FR="manger"+FLX=SIP1_G+FACT+pex+1
mikuchinkichik,mikuy,V+FR="manger"+FLX=SIP1_G+FACT+p+2
mikuchinku,mikuy,V+FR="manger"+FLX=SIP1_G+FACT+p+3
mikuchkani,mikuy,V+FR="manger"+FLX=SIP1_G+PROG+s+1
mikuchkanki,mikuy,V+FR="manger"+FLX=SIP1_G+PROG+s+2
mikuchkan,mikuy,V+FR="manger"+FLX=SIP1_G+PROG+s+3
mikuchkanchik,mikuy,V+FR="manger"+FLX=SIP1_G+PROG+pin+1
mikuchkaniku,mikuy,V+FR="manger"+FLX=SIP1_G+PROG+pex+1
mikuchkankichik,mikuy,V+FR="manger"+FLX=SIP1_G+PROG+p+2
mikuchkanku,mikuy,V+FR="manger"+FLX=SIP1_G+PROG+p+3
mikuikachani,mikuy,V+FR="manger"+FLX=SIP1_G+DISP+s+1
mikuikachanki,mikuy,V+FR="manger"+FLX=SIP1_G+DISP+s+2
mikuikachan,mikuy,V+FR="manger"+FLX=SIP1_G+DISP+s+3
mikuikachanchik,mikuy,V+FR="manger"+FLX=SIP1_G+DISP+pin+1
mikuikachaniku,mikuy,V+FR="manger"+FLX=SIP1_G+DISP+pex+1
mikuikachankichik,mikuy,V+FR="manger"+FLX=SIP1_G+DISP+p+2
mikuikachanku,mikuy,V+FR="manger"+FLX=SIP1_G+DISP+p+3
mikuikachini,mikuy,V+FR="manger"+FLX=SIP1_G+POLI+s+1
mikuikachinki,mikuy,V+FR="manger"+FLX=SIP1_G+POLI+s+2
mikuikachin,mikuy,V+FR="manger"+FLX=SIP1_G+POLI+s+3
mikuikachinchik,mikuy,V+FR="manger"+FLX=SIP1_G+POLI+pin+1
mikuikachiniku,mikuy,V+FR="manger"+FLX=SIP1_G+POLI+pex+1
mikuikachinkichik,mikuy,V+FR="manger"+FLX=SIP1_G+POLI+p+2
mikuikachinku,mikuy,V+FR="manger"+FLX=SIP1_G+POLI+p+3
```

Figure 2. Sample of the 5175 inflected forms of the verb mikuy (to eat)

One strategy of Quechua for enlarging its lexicon of verbs is the derivation. In fact, when we interpose one of the 26 inf-interposed suffixes[11] between the verbal lemma and the infinitive particle *y* we get 28840 (26*1400) new verbs. Which not only produce the respective nominalizations (adding the suffixes i, q, na) but also will imply the generation of plenty of MWU.

4 Identifying MWU: Outputs of Nooj Grammar Queries

As we can remark in the following examples of MWU, the collocation of two PoS, inflected or not, gives us a form which does not necessarily remain in the semantic field of either of the components.

yana uma (lit. black(A) head(N)) becomes traitor(A).

sunqu suwa (lit. heart(N) becomes thief(A)) a flirt (A)

piki piki (lit. flea(N) flea) becomes very fast (A)

[7] Duran, M. (2015) "The annotation of compound suffixation structure of Quechan verbs". In Proceedings of the 2015 International NooJ Conference, National Academy of sciences. Minsk. Belarus. (To appear)

[8] The other key structure is for the future tense
FUT = <B>(*saq*/F+s+1 | *nki*/F+s+2 | *nqa*/F+s+3 | *saqku*/F+pex+1 | *sunchik*/F+Pin+1 | *nkichik*/F+p+2 | *nqaku*/F+p+3);

[9] IPS = *{ chi, chka, ikacha, ikachi, ikamu, ikapu, ikari, iku, isi, kacha, kamu, kapu, ku, lla, mpu, mu, na, naya, pa, paya, pti, pu, ra, raya, ri, rpari, rqa, rqu, ru, spa, sqa, tamu, wa }* (33)

[10] PPS={- ch, chaá, chik,… má, man, m, ña, pas, puni, qa, raq, s, taq, yá }(17)

[11] inf_IPS =IPS-{na, ra, rqa, spa, stin, wa, ru};

qallu qallu (lit. tongue (N) tongue(N)) becomes a parasite of the lever (N)
kachi kachi (lit. salt(N) salt(N)) means dragonfly (N)

Applying some Nooj grammars like the one in Figure 3. on our corpus of around 80 000 tokens, we can obtain lists of potential noun_noun MWU like it appears in the central column of Figure 4.

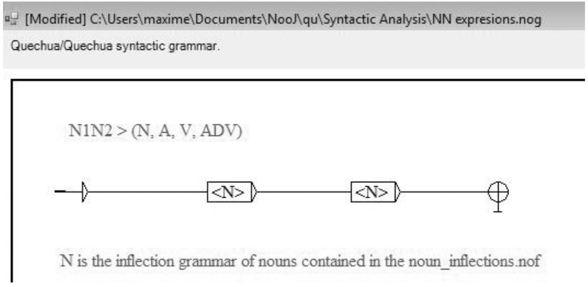
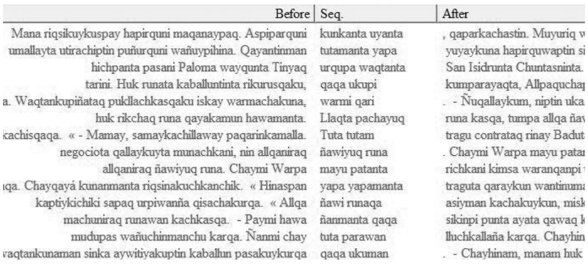

[Modified] C:\Users\maxime\Documents\NooJ\qu\Syntactic Analysis\NN expresions.nog
Quechua/Quechua syntactic grammar.
N1N2 > (N, A, V, ADV)
<N>
<N>
N is the inflection grammar of nouns contained in the noun_inflections.nof

Figure 3. This N_N grammar identifies the noun_noun MWU in a text.



Before	Seq.	After
Mana riqsikuykuspay hapirquni maqanaypaq. Aspiparquni	kunkanta uyanta	, qaparkachastin. Muyuriq w
umallayta utirachiptin puñurquni wañuypihina. Qayantinman	tutamanta yapa	yuyaykuna hapirquwaptin si
hichpanta pasani Paloma wayqunta Tinyaq	urqupa waqtanta	San Isidrunta Chuntasninta.
tarini. Huk runata kaballuntinta rikurusqaku,	qaqa ukupi	kumparayaqta, Allpaqucha
a. Waqtankupiñataq pukllachkasqaku iskay warmachakuna,	warmi qari	. - Ñuqallaykum, niptin uka
huk rikchaq runa qayakamun hawamanta.	Llaqta pachayuq	runa kasqa, tumpa allqa ñav
kachisqaqa. « - Mamay, samaykachillaway paqarinkamalla.	Tuta tutam	tragu contrataq rinay Badut
negociota qallaykuyta munachkani, nin allqaniraq	ñawiyuq runa	. Chaymi Warpa mayu patar
allqaniraq ñawiyuq runa. Chaymi Warpa	mayu patanta	richkani kimsa waranqanpi
ıqa. Chayqayá kunanmanta riqsinakuchkanchik. « Hinaspan	yapa yapamanta	traguta qaraykun wantinuma
kaptiykichiki sapaq urpiwanña qisachakurqa. « Allqa	ñawi runaqa	asiyman kachakuykun, misk
machuniraq runawan kachkasqa. - Paymi hawa	ñanmanta qaqa	sikinpi punta ayata qawaq k
mudupas wañuchinmanchu karqa. Ñanmi chay	tuta parawan	lluchkallaña karqa. Chayhin
vaqtankunaman sinka aywitiyakuptin kaballun pasakuykurqa	qaqa ukuman	. - Chayhinam, manam huk

Figure 4. Output of the Nooj grammar which retrieves N_N multiword units in the corpus.

How does it work? Let us take the third line form: *urqupa waqtanta* (by the mountain's side); Nooj proposes this collocation because in one of the dictionaries of inflected forms, among the thousands of inflections of the noun (N) *urqu*: the mountain, it has found *urqupa*: belonging to the mountain, and also the inflected form of the adjective (A) *waqta*: side, it has found *waqtanta*:by the side of. For a N_A collocation this fits with the rule then it proposes it.
But, not all of these outputs are actually valid MWU. For instance in line 12th Nooj wrongly proposes the form (N1_N2) *ñanmanta qaqa*: (lit. rock from the path) as a MWU. For this case of N1_N2 multi word unit, the rule says that only the second component may be inflected not the first one, then it should not be proposed as a MWU and yet it is.
Hence we need to introduce this rule as a filter. Once this filter is in the program the (N1_N2) *ñanmanta qaqa* is not proposed as a MWU anymore. In a similar way the collocation *qara-n uya-ri* will now be rejected by NooJ as a candidate to be a MWU because it finds that the first component *qara-n*: its leather, its skin, is an inflected form; whereas it accepts as a valid inflection: *qara uyan-wan* : behaving as a rascal.
Thus to avoid the over production and ambiguities, it is necessary to introduce disambiguation grammars containing filters like the one in Figure 5. based in the Quechua morpho-syntaxis.

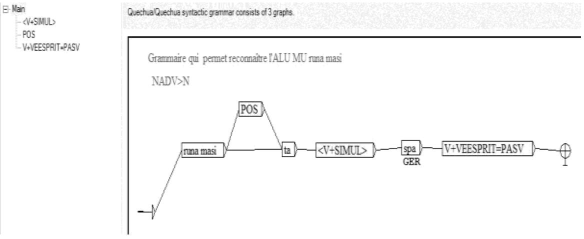

Main
<V+SIMUL>
POS
V+VEESPRIT=PASV
Quechua/Quechua syntactic grammar consists of 3 graphs.
Grammaire qui permet reconnaitre l'ALU MU runa masi
NADV>N
POS
runa masi
ta
<V+SIMUL>
spa
GER
V+VEESPRIT=PASV

Figure 5. A disambiguation grammar for N-ADV collocations.

An important resource for handling MWU is to build manually a dictionary of MWU:a Lexicon-grammar of MWU, developed following a similar framework as M. Gross (1982) . To do it we have programmed 12 syntactic NooJ grammars like in Figure 3. one for each class of collocation listed in the introduction, and 12 corresponding disambiguation ones.
We have arrived to manually gather more than 1000 MWU of two components, in a lexicon named QU-MWU. It has the form of an electronic grammar containing their PoS categories, their French and Spanish translations, and the accompanying flexional grammar as shown in Figure 6.

kuyaypaq kaq, A+MWE+UNAMB +FR="carismatique" +SP=" carismático"+FLX= A_G_1
runa kay, N+MWE+UNAMB +FR="l'être humain"+SP= "ser humano" +FLX= N_G_1
qepa punchau, A+MWE+UNAMB +FR=" le passé" +SP="el pasado" +FLX= A_G_1
kuchun kuchun, A+MWE+UNAMB +FR=" tout les coins « +SP=« de rincon en rincon » +FLX= N_G_1
wallpa waqayta, N+MWE+UNAMB +FR=" à l'aube «+SP=« al amanecer"+FLX= N_G_1
tawa chaki, A+MWE+UNAMB +FR=" cuadrupède, bête «+SP=« cuadrúpedo, bestia" +FLX= A_G_1
chulla ñawi, A+MWE+UNAMB +FR=" bigorne «+SP=« bisco" +FLX= N_G_1
sunqu suwa, A+MWE+UNAMB +FR=" voleur des coeurs «+SP=« ladron de corazones » +FLX= A_G_1
qaqa uku, N+MWE+UNAMB +FR=" abïme «+SP=« abismo » +FLX= N_G_1

Figure 6. A sample of the tri-lingual Qu_MWU lexicon .

We expect that it will serve us as a linguistic resource in the recognition, the annotation of

MWU and in the further project of machine translation technology.

5 Morphology of Quechua MWU Sentences

Many MWU are frozen units, but many more can be transformed as we have seen in the introduction, by inflexions applied to one of the components without changing the semantic value of the MWU. For instance let us take the N_N MWU *piki piki* > rapidly (ADV) (*piki* (N): flea):

Pablo llamkan **piki piki**-cha Pablo Works rapidly and in a short time

Pablo llamkan **piki piki**-lla Pablo Works rapidly with care

Pablo llamkan **piki piki**-lla-ña Pablo Works rapidly already with care

Where the second noun: piki (N) appears in several inflected forms.

The next example concerns a N_ADV MWU: ***runa masi*** > human kindness (A) (*runa* (N) *masi* (ADV): similar)

runa masi*-nchik* our fellow human

runa masi*-nchik-ta qawaspa kusikuni*

I am happy seeing our fellow human

runa masi*-nchik-wan kusikuni*

I am happy with our fellow human

runa masi*-nchik-raiku kusikuni*

I am happy for our fellow human sake

Here also it is the second component: *masi* (ADV) which is inflected. The same grammar in Figure 3. will generate a large number of trans-formation of the form *runa masi* (POS, ta) V1+SIMUL V2+PR

We may propose the hypothesis that:

- if C1_C2 is a MWU, where C1 is not a verb and can be inflected, it is the second component (C2) that will bear the inflections.

- if the first component C1 is a verb, the MWU may appear with C1 or C2 inflected

Example: In the MWU *kuaypaq kaq* (a nice person) both components may be inflected kuyaypaq*(mi, cha, chus?, si,…) ka (q, nki,n, ptin):kuyaypaqmi kaq* (he used to be nice).

I have not yet arrived to program the automatic generation of all the valid transformation of this class of MWU.

6 On the Syntactic Grammar for Para-phrase Generation involving MWU

The Syntactic Grammar which generates paraphrases/transformations of phrases containing MWU takes into account the restrictions on the applicability of transformations given by the inflectional and derivational grammars of its components.

Quechua does not have prepositions neither conjunctions, which may help in the generation of paraphrases, it is the set of suffixes imbedded in the inflections that accomplish these roles as we can see in the MWU *runa masi* fellow human:

Pablo riman Inesta ***runa masi****-n-man hina*

Pablo talks to Ines as if he was his fellow human.

Some of its corresponding paraphrases are:

runa masi*-n-man hina Pablo riman Inesta*

As if he was his fellow human, Pablo talks to Ines

runa masi*-n-man hina Pablo Inesta riman*

As if he was his fellow human, Pablo talks to Ines

Ines-ta Pablo riman ***runa masi****-n-man hina*

Pablo talks to Ines as if he was his fellow human

Pablo-m Ines-ta ***runa masi****-n-man hina* It is Pablo who talks to Ines as if he was his fellow human

runa masi-n-man hina Pablo Inesta riman

As if he was his fellow human Pablo, talks to Ines

A phrase like

Rosam Pablopa umanta quñichin (Rose has turned Pablo's head)[12] has been analyzed within the model of M. Gross (1982).

It fallows the structure: **N1(m)_ N2(pa) C1V,** where N1 and N2 represent the free constituents and V, C1 indicate the frozen parts, -*m* and –*pa* are nominal suffixes.

To generate all of the possible paraphrases we have programmed the graphical grammar appearing in Figure 7. This grammar is formed of 12 embedded grammars, it allows the generation/annotation of 9 elementary paraphrases and at least 86 possible combinations of paraphrases.

All the agreement constraints are necessary in order to generate only grammatical sentences. If they are not set, NooJ will produce ungrammatical results. After the syntactic grammar is built, it is possible to generate the paraphrases of a given QU-MWU by right clicking on the syntactic grammar, selecting the Produce Paraphrases function and entering the QU-MWU sentences.

NooJ will produce 86 paraphrases like:

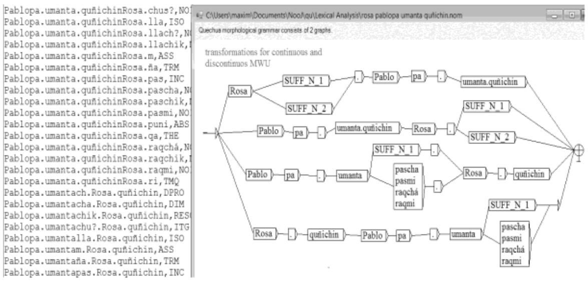
Rosam Pablopa umanta quñichin

Rosam Pablopa umantaja quñichin

Pablopa umantam quñichin Rosa

[12] This example is inspired in that of S. Vietri (2012)

Pablopatam umanta quñichin Rosa, …
Rosam Pablopa umanta quñi-rqa-chin (quñi-ra-chin, quñi-rqa-chin, quñi-paya-chin, quñi-mpu-chin, quñi-pa-chin, quñi-ri-chin, quñi-isi-chin, quñi-naya-chin…)



```
Pablopa.umanta.quñichinRosa.chus?,NO
Pablopa.umanta.quñichinRosa.lla,ISO
Pablopa.umanta.quñichinRosa.llach?,N
Pablopa.umanta.quñichinRosa.llachik,
Pablopa.umanta.quñichinRosa.m,ASS
Pablopa.umanta.quñichinRosa.ña,TRM
Pablopa.umanta.quñichinRosa.pas,INC
Pablopa.umanta.quñichinRosa.pascha,N
Pablopa.umanta.quñichinRosa.paschik,
Pablopa.umanta.quñichinRosa.pasmi,NO
Pablopa.umanta.quñichinRosa.puni,ABS
Pablopa.umanta.quñichinRosa.qa,THE
Pablopa.umanta.quñichinRosa.raqchá,N
Pablopa.umanta.quñichinRosa.raqchik,
Pablopa.umanta.quñichinRosa.raqmi,NO
Pablopa.umanta.quñichinRosa.ri,TMQ
Pablopa.umantach.Rosa.quñichin,DPRO
Pablopa.umantacha.Rosa.quñichin,DIM
Pablopa.umantachik.Rosa.quñichin,RES
Pablopa.umantachu?.Rosa.quñichin,ITG
Pablopa.umantalla.Rosa.quñichin,ISO
Pablopa.umantam.Rosa.quñichin,ASS
Pablopa.umantaña.Rosa.quñichin,TRM
Pablopa.umantapas.Rosa.quñichin,INC
```

C:\Users\maxim\Documents\NooJ\qu\Lexical Analysis\rosa pablopa umanta quñichin.nom

Quechua morphological grammar consists of 2 graphs.

Figure 7. Sample of paraphrases obtained with the paraphrase generator grammar.

7 Conclusion and Perspectives

We have presented some morpho-syntactic grammars programmed in the linguistic platform NooJ, which allow us to identify and disambiguate two-components MWU in a text.

Using these grammars, we have shown how we have constituted a lexicon of more than 1000 two-component MWU coming from the written corpus.

Since Quechua remains dominantly an oral language, in case of coming projects, we have underlined the need of deploying significant efforts to gather manually more MWU coming from field work. We have already started to gather a MWU lexicon of more than two components.

We have presented one graphical morpho-syntactic grammar to generate paraphrases of a phrase that contains two-components MWU.

References

César Guardia Mayorga. 1973. *Gramatica Kechwa*, Ediciones los Andes, Lima. Peru.

Cesar Itier. 2011. *Dictionaire Quechua-Français*, Paris. L'Asiathèque, Paris.

Diego Gonzales Holguin, 1608. *Vocabulaire de la Lengua General de todo el Perú llamada Lengua Qquichua o del Inca*, Universidad Nacional Mayor de San Marcos 1952), Lima.

Maurice Gross. 1982. *Une classification des phrases "figées" du français.* Revue Québécoise de Linguistique 11.2, Montreal: UQAM.

Max Silberztein. 2003. *NooJ Manual*. Available at: htpp://www.nooj4nlp.net (220 pages updated regularly).

Max Silberztein. 2010. *Syntactic parsing with NooJ*, in Proceedings of the NooJ 2009 International Conference and Workshop, Sfax: Centre de Publication Universitaire: 177-190.

Max Silberztein. 2015. *La formalisation des langues, l'approche de NooJ*. ISTE Editions (425 pages), London.

Maximiliano Duran. 2009. *Diccionario Quechua-Castellano*. Editions HC. Paris.

Pedro Clemente Perroud, 1972. *Gramatica Quechwa Dialecto de Ayacucho*. Lima. 3ª Edicion.

Simonetta Vietri. 2012. *Transformations and Frozen Sentences*, in Proceedings of the 2011 International Nooj Conference. Cambridge Scholars Publishing:166-180.

Populating a Lexicon with Multiword Expressions in view of Machine Translation

Voula Giouli
Institute for Language and Speech Processing, Athena RIC
National and Kapodistrian University of Athens, Greece
voula@ilsp.gr

Abstract

We herby present work aimed at populating a computational semantic lexicon of Modern Greek with Multiword Expressions (MWEs) and their encoding in it. The lexicon is organized as a conceptual dictionary that would be applicable for a range of Natural Language Processing (NLP) applications. Entries in the lexicon are organized around specific, predefined domains or semantic fields; rich lexical, syntactic and semantic information is provided for every lexical lexicon entry, and translational equivalences in English (EN) are added. Single- and multiword entries are mapped onto sets of concepts that are specific to the domains or semantic fields at hand. In this view, the Language Resource (LR) developed caters for cross-lingual and interlingual alignments that would be valuable for Machine Translation.

1 Introduction

Multiword Expressions (MWEs) are lexical items characterized by lexical, syntactic, semantic, pragmatic or statistical idiosyncrasies. And although they may be defined on the basis of sound linguistic criteria, they appear in a variety of configurations and a continuum of *compositionality,* which ranges from expressions that are very analysable to others that are partially analysable or ultimately non-analysable (Nunberg et al. 1994). In this respect, they pose a challenge to the automatic processing of human languages. In recent years, there is a growing interest within the NLP community in the identification of MWEs and their robust treatment, as this seems to improve parsing accuracy (Nivre and Nilsson, 2004; Arun and Keller, 2005) or MT quality (Ren et al., 2009; Carpuat and Diab 2010). In this paper, we present work aimed at the development of a large-scale LR that encompasses MWEs. Existing typologies have been employed as the basis for their efficient representation; these typologies were further enriched and extended with new information in view of MT.

The paper is structured as follows. The basic motivation and scope of our work is outlined in section 2; section 3 gives an overview of background work on LRs that encompass MWEs. The data-driven methodology adopted in view of populating the LR with MWEs, namely Support Verb Constructions and nominal MWEs, is described in section 4 focusing on the datasets used for extracting various types of MWEs, and for modelling their underlying structure in view of corpus evidence. Section 5 gives an account of the semi-automatic extraction of candidate nominal MWEs from textual data, and the work performed for the identification of translational equivalents in aligned texts; the annotation applied to the MWEs is also described here. Our proposal towards enriching the encoding scheme is presented in section 6, while section 7 outlines our conclusions and prospects for future research.

2 Motivation and Scope

The purpose of the work presented here is twofold; on the one hand, we aimed at the identification and semi-automatic extraction of Greek MWEs in the selected *domains/semantic fields* from textual data. More precisely, the domains *administration*, *education*, *health, sports* and *travel* were targeted along wth the semantic fields *emotion* and *cognition.* Moreover, a number of features or properties of the selected MWEs were also identified as important in the automatic treatment of MWEs and encoded in the lexicon. In this respect, one major challenge was the representation of MWEs in such a way that might be useful for prospect applications with MT in perspective.

From another perspective, the study aimed at the identification of cross-lingual correspondences between the EL and EN expressions. This was achieved in two ways: (a) explicitly, by means of their translational equivalents in English (where applicable) and/or (b) implicitly, providing their lexical semantic relations (i.e, synonyms, antonyms).

The focus will be on the following aspects: (a) identification and manual extraction of MWEs in the selected *domains/semantic fields* at the monolingual level from a set of parallel domain-specific EL-EN corpora, (b) alignment of MWEs, and (c) encoding of MWEs in the database.

3 Background

The lexicon presented here builds on an existing conceptual dictionary developed for the Greek Language (Markantonatou & Fotopoulou, 2007). Lexical entries are represented in the dictionary modeling the notion of the linguistic SIGN and its two inseparable facets, namely, the SIGNIFIER and the SIGNIFIED. The final resource forms a linguistic ontology in which words (word forms) are instances in the SIGNIFIER class; these are further specified for (a) features pertaining to lexical semantic relations (i.e, synonymy, antonymy); (b) lexical relations such as word families, allomorphs, syntactic variants etc.; and (c) morphosyntactic properties (PoS, gender, declension, argument structure, word specific information etc.). Values for these features are assigned to both single- and multi-word entries in the lexicon. Similarly, word meanings are instances in the SIGNIFIED class. Each instance in the SIGNIFIER class is mapped onto a concept, the latter represented as an instance in the SIGNIFIED class. In this context, MWEs are represented in the SIGNIFIER class (and are coupled with rich linguistic information pertaining to the lexical, syntactic and semantic levels); mappings to the relevant senses are provided thereof. An encoding schema that applies to verbal and nominal MWEs specifically has been proposed (Fotopoulou et al, 2014). At the level of concepts, words in the lexicon (their senses) are also organised in semantic fields, defined as groups of lexemes which share a common sense, i.e., the semantic field o emotion, cognition, travel, health, etc.

In terms of surface structure, the typology of verbal MWEs shares common characteristics with similar efforts for other languages. In the computational dictionary for the Dutch language DuELME (Gregoire, 2010) MWEs are grouped according to their syntactic pattern. Each group is represented by a pattern identifier which is documented in a detailed pattern description. This description includes: (a) the syntactic category of the head of the expression, (b) its complements, (c) a description of the internal structure of the complements, and (d) morphosyntactic information of the individual components.

4 Methodology

A data-driven approach has been adopted both in the acquisition and description of MWEs. As a first step, initial lists of candidate MWEs were identified on the basis of corpus evidence. To this end, the Hellenic National Corpus (HNC), a large (tagged and lemmatised) reference corpus for the Greek language (Hatzigeorgiou et al, 2000) was employed. The corpus, which amounts to circa 47M words, comprises written texts of a broad range of text types / genres and topics from various sources, thus representing the synchronic usage of standard Modern Greek. Through the web interface[1], HNC can be queried for wordforms, lemmas and morphosyntactic (Part-of-Speech, POS) tags or any combination thereof. Results are visualised as concordances and frequency information (statistics). A major functionality of its web interface, allows for the HNC to be searched either as a whole, or on the basis of sub-corpora defined according to classification and annotation parameters that accompany each text, thus creating sub-corpora of a specific author, or belonging to a specific genre, text type, domain etc. In this line, extensive searches were performed on selected predefined sub-corpora of the whole corpus that pertain to the domains at hand. Existing linguistic typologies of Greek verbal MWEs (Fotopoulou, 1993, Mini, 2009) and of nominal MWEs (Anastasiadis, 1986), which depict the lexical and syntactic configurations involved, were taken into account in this respect. This process

[1] http://hnc.ilsp.gr

was particularly useful and efficient for the identification of candidate Support Verb Constructions (SVCs). The following support verbs were considered: *βάζω* (=put), *δίνω* (=give), *κάνω* (=make), *λαμβάνω* (=get, in its formal register), *παίρνω* (=get), and *ρίχνω* (=drop). The resulting structures were further manually checked for the exclusion of collocations or structures which do not fall in the SVCs. This process resulted in the identification of 1692 SCVs for inclusion in the lexicon.

The afore-mentioned procedure yielded MWEs that were classified in various domains/subject fields. At the next stage, the acquisition of in-domain MWEs was also addressed. To this end, a suite of specialized bilingual corpora in Greek (EL) and English (EN) were selected manually from various sources over the web. More specifically, the bilingual corpus comprises texts that adhere to predefined domains, namely *administration*, *education*, *health*, *sports* and *travel*, targeted at during the dictionary development process. Depending on availability of suitable data, the resulting corpus comprises sub-corpora that are either parallel (*administrative*, *education*, *travel*) or comparable (*health*, *sports*). Comparable corpora were collected from Wikipedia, and comparability criteria (Skandina et al., 2010) in terms of size were taken into account. In-domain nominal MWEs were extracted semi-automatically from these corpora (section 5.1 below).

Moreover, to cater for the acquisition of lexical data that pertain to the semantic fields of *emotion* and *cognition*, EL texts from online blogs and forums were collected. The latter is considered as user-generated content, conveying the emotions and opinions of users with respect to certain subject matters, products, etc. Moreover, they depict a more informal, everyday language. These corpora were manually annotated for the identification of MWEs and their properties.

The final corpus amounts to c. ~220K tokens; its distribution with respect to the domains or subject fields as described above is presented in Table 1 below.

Domain	**Languages-Corpus type**	**Tokens**
Administrative	EL-EN, parallel	25777
Education	EL-EN, parallel	32219
Health	EL-EN, comparable	9411
Sports	EL-EN, comparable	5128
Travel	EL-EN, parallel	33478
Opinionated Articles	EL	32989
Blogs	EL	38619
Forums	EL	42074
Total		**219695**

Table 1. Corpus Distribution

5 Corpus Processing and Annotation

Corpus processing was performed on the textual data along two axes: monolingual and bi-lingual. At the monolingual level, processing was aimed at boosting the process of extraction and selection of MWEs that pertain to the designated semantic fields or domains. In this sense, identification of candidate MWEs was treated as a term extraction task; the focus, however, was placed on multi-word candidates rather than single-word ones. Moreover, shallow processing is a prerequisite for

At the bilingual level, processing was aimed at the alignment of parallel texts at the sentence level, and the identification of the translational equivalent of each MWE.

Moreover, to better account for the identification and efficient encoding of linguistic properties of the MWEs identified, the annotation of Greek corpora was also performed. In the next sections, we will elaborate further on the tasks performed.

5.1 Semi-automatic extraction of in-domain MWEs

The task of in-domain MWE extraction was viewed as a term extraction one; yet the focus is placed on multi-word expressions rather than single word terms. Defined as a lexical unit comprising one or more words, the notion of *term* is used to represent a concept inside a domain. In this respect, an existing pipeline of shallow processing tools for the Greek language (Papageorgiou et al, 2003) was employed to process EL texts that pertain to the specific domains (administrative, education, health, sports, travel). Processing involves tokenization, sentence splitting, Part-Of-Speech (POS) tagging, and lemmatization; at the final stage, candidate terms were identified in the texts are then statistically evaluated with an aim to skim valid domain terms and lessen the over-generation effect caused by pattern grammars. To this end, a confidence score is assigned to each candidate term.

Manual validation of the output involved discarding single-word candidates and n-grams which are not MWEs. The system favours sequences of bi-grams assigning them a higher confidence score compared to tri-grams or unigrams. The results are depicted in Table 2 below.

Domain	TE-all	TE-MWEs	MWEs-correct
Administrative	1153	987	851
Education	1570	1107	996
Health	2984	1205	1087
Sports	809	186	160
Travel	1287	608	580
Total	7803	4093	3675

Table 2. Results

5.2 Acquisition of MWE translations: alignment of parallel texts

Finally, to facilitate the acquisition of translational equivalents of MWEs, the EL (source) texts were aligned with their translations in EN. Alignment at the sentence level was also performed semi-automatically on the bilingual parallel sub-corpora (i.e., texts that pertain to the administrative, education and travel domains) using UNITEX platform (Paumier 2013) and its built-in functionality XAlign.

Pattern matching queries were applied on the Source EL texts and the alignments of specific structures were returned by the corpus management tool. To further exploit parallel texts, and the alignments acquired, the tool facilitates retrieval of sentences aligned with matched sentences. This means that once a set of segments (sentences) is matched against a query, Unitex filters all the remaining unmatched segments. So, it is easy to lookup for an expression in one text and to find the corresponding sentences in the other.

5.3 MWE annotation in Greek texts

To better account for the lexical, syntactic and semantic properties of the MWEs identified in the corpora, their annotation in the EL texts was in order. The resulting corpus facilitates modeling the underlying structures in light of corpus evidence. From another perspective, the resource might also be used as a testbed for guiding the development and evaluation of a tool for the identification of candidate MWEs. Annotation was performed on pre-processed (i.e., tokenized, POS-tagged and lemmatized) text. The annotation schema adopted was aimed at the following: (a) identification of MWE extent, (b) MWE classification with respect to grammatical category (POS), and (c) fixedness information.

Defining the grammatical category of MWEs was not always straightforward. This was especially true for MWEs that were ultimately classified as having an adjectival usage, as depicted in the following examples:

(1) *πυρ και μανία*
pir ke mania
lit. fire and fury
very angryAJ)

(2) *εκτός εαυτού*
ektos eaftu
*lit.*outside oneselfSG.GEN
very angryAJ

However, the specifications set for the tasks of identification and classification of MWEs make extensive use of linguistic criteria (semantic, lexical and morphosyntactic) (Gross, 1982, 1988; Lamiroy, 2003), namely:

- non-compositionality: i.e., the meaning of the expression cannot be computed from the meanings of its constituents and the rules used to combine them;
- non-substitutability: at least one of the expression constituents does not enter in alternations at the paradigmatic axis
- non-modifiability: MWEs are syntactically rigid structures, in that there are constraints concerning modification, transformations, etc.

These criteria, however, do not apply in all cases in a uniform way. As a matter of fact, fixed expressions appear in a continuum of compositionality, which ranges from expressions that are very analyzable to others that are partially analyzable or ultimately non-analyzable The variability attested brings about the notion 'degree of fixedness' (Gross, 1996). The kind and degree of fixedness result in the classification of these expressions as *fixed, semi-fixed, syntactically flexible* or *collocations* (Sag et al, 2002).

A typology of Greek verbal MWEs has been defined in (Fotopoulou, 1993, Mini, 2009) (NP V NP1 NP2…) and of nominal MWEs in (Anastasiadis, 1986) (AJ N, NN…) on the basis of the lexical and syntactic configurations involved. This typology has been extended with new classes in light of the annotated material. Examples of the new MWE classes are depicted in Table 3 below.

Tags	Tags-new	POS	Example

Aj No		No	πρωτοβάθμια εκπαίδευση (=primary education)
Aj Aj No	Aj MWE-No	No	ανώτατο εκπαιδευτικό ίδρυμα (=higher education institution)
Aj No No-ge		No	
No Aj No-ge		No	κέντρο επαγγελματικής κατάρτισης
No No-ge		No	φακός επαφής (contact lenses)
	No No-ge No-ge	No	έκδοση άδειας λειτουργίας
	No AtDf-ge No-ge	No	συνήγορος του πολίτη
No No		No	νόμος πλαίσιο (law-frame)
	Ad No-ge	Aj	έξω φρενών (*lit*., "out of-mind", =furious)
	No CONJ No	Aj	πυρ και μανία (*lit*., :fire and fury" =very furious)
	No No-ge	Aj	χάρμα οφθαλμών (=beautiful)
	No PP	Aj	κεραυνός εν αιθρία (=sudden)

Table 3. Example of new MWE types

6 Extending the encoding scheme

At the last stage, encoding of the MWEs and their classification according to the observed properties was performed. The initial encoding scheme has already been extensively described in Fotopoulou et al (2014). We have opted for an approach to MWE representation that builds on rich linguistic knowledge. The linguistic classifications adopted deal with morphology, syntax, and semantics interface aspects. Thus, a lexicon – grammar representation of MWEs has been constructed by encoding key morpho-syntactic and semantic information. Morphosyntactic properties and selectional preferences account better for the idiosyncrasies of the Greek language, as for example word order and gaps attested in running text. In the remaining, the new additions and/or modifications to the schema will be discussed.

6.1 MWE lemma form

Apart from the surface form, lemma information and morphosyntactic features that were available from the processed data were encoded in the lexicon.

<MWE id="mwe130" name="σύμβαση ανταλλαγής κινδύνου αθέτησης" lemma="σύμβαση ανταλλαγή κίνδυνος αθέτηση" POS="Vb" POS-C="NoNm NoGe NoGe NoGe " /MWE>

<MWE id="mwe235" name="μου τη δίνει" lemma="εγώ του δίνω" POS="Vb" POS-C="PnPeWeGe PnPeWeAc Vb" /MWE>

6.2 MWE fixedness

As it has been said, multi-word expressions often occur in texts in various configurations. The encoding of fixed and non-fixed constituents provides, therefore, extra information for the identification of expressions in texts. Moreover, the identification of MWEs as collocations entails a relatively loose fixedness, allowing, thus, for gaps and discontinuities as shown in (2):

(3) *Το κόμμα έχει αριθμό* υποψηφίων-*ρεκόρ*
to koma echi ariθmo ipopsifion-rekor
*lit.*The party has number of-candidatesPL.GEN record
the political party has many candidates

Word order is not always fixed in certain expressions, and a loose word order is occasionally allowed:

(4) *καρδιακός φίλος / φίλος καρδιακός*
karδiakos filos / filos karδiakos
*lit.*heart friend / friend heart
good friend / friend good

Moreover, modification and certain structural transformations are in cases allowed as shown in examples (5) and (6) below:

(5) *κάνω την αρχή*
kano tin archi
lit. make theDEF start
make a start

(6) *κάνω μια*INDEF **νέα**AJ *αρχή*
kano mia nea archi
make a new start

6.3 Subcategorisation information and semantic role labeling

In order to ensure a uniform representation of both single- and multi-word entries in terms of the, non-fixed arguments (i.e., complements) of verbal MWEs were encoded formally. Syntactic alternations that are relevant are also provided for as features connected to the arguments at hand:

(7) *κάνω αναφορά* ***Arg0****=NP* ***Arg1****=PPσε*

7 Conclusion and future research

We have presented work aimed at populating a conceptual lexicon of modern Greek with MWEs that pertain to specific domains or semantic fields semi-automatically from corpora. To make the resource applicable for NLP applications a number of properties were encoded as features in the lexicon, as for example, surfuce structure of the MWE, lemma information of all the constituents, subcategorisation information, etc. Future research involves the systematization of the properties attested in the data collected and the annotated corpora, and the definition of an extended typology across grammatical categories in view of developing a rule-based system that recognizes MWEs in running text.

Acknowledgements

The research leading to these results has partly received funding from POLYTROPON project (KRIPIS-GSRT, MIS: 448306).

References

Anastasiadi-Simeonidou, A. 1986. *I Neologia stin Koini Nea Elliniki*. Aristotle University of Thessaloniki, Greece. (In Greek).

Arun, A., and F. Keller. 2005. Lexicalisation in crosslinguistic probablisitic parsing: The case of French. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 306–313. Ann Arbor, MI.

Carpuat, M., and Diab, M. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. Human Language Technologies: In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10). Association for Computational Linguistics, Stroudsburg,* PA, USA, 242-245.

Fotopoulou, A., Markantonatou, S. and Giouli, V. 2014. Encoding MWEs in a conceptual lexicon. In *proceedings of the 10th EACL Workshop on Multiword Expressions (MWE 2014)*. Gothenburg, Sweden. April 26–27, 2014.

Fotopoulou, A. 1993. *Une Classification des Phrases à Compléments Figés en Grec Moderne*. Doctoral Thesis, Universite Paris VIII.

Gregoire, N. 2010. DuELME: a Dutch electronic lexicon of multiword expressions. In *Language Resources & Evaluation (2010) 44:23–39*.

Gross, M. 1982. Une classification des phrases figées du français. *Revue Québécoise de Linguistique 11 (2), 151-185.*

Gross, M. 1988. Les limites de la phrase figée. *Langage 90: 7-23.*

Lamiroy, B. 2003. Les notions linguistiques de figement et de contrainte. *Lingvisticae Investigationes* 26:1, 53-66, Amsterdam/Philadelphia: John Benjamins.

Markantonatou, S., and Fotopoulou, A. 2007. The tool “Ekfrasi”. In *Proceedings of the 8th International Conference on Greek Linguistics, The Lexicography Workshop*. Ioannina, Greece.

Mini, M. 2009. *Linguistic and Psycholinguistic Study of Fixed Verbal Expressions with Fixed Subject in Modern Greek: A Morphosyntactic Analysis, Lexicosemantic Gradation and Processing by Elementary School Children*. Unpublished doctoral dissertation. University of Patras.

Nivre, J. and Nilsson, J. 2004. Multiword units in syntactic parsing. In *Proceedings of the Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*.

Nunberg, G., Sag I., and Wasow, T. 1994. Idioms. *Language 70*, pp. 491-538.

Papageorgiou, H., Prokopidis, P., Giouli, V., Demiros, I., Konstantinidis, A., Piperidis, S., 2002. Multi-level, XML - based Corpus Annotation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Paumier, S. 2003. *UNITEX User Manual*.

Ren, Z., Lü, Y., Cao, J., Liu, Q., and Huang, Y. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE '09).* Association for Computational Linguistics, Stroudsburg, PA, USA, 47-54.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. 2002. Multiword Expressions: A Pain in the Neck for NLP. In A. F. Gelbukh, ed. 2002. *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '02)*. Springer-Verlag, London, UK, UK, 1-15.

Skadina, I., Aker, A., Giouli, V., Tufis, D. Gaizauskas, R., Mierina, M., and Mastropavlos, N. 2010. Collection of comparable corpora for under-resourced languages. In *Proceedings of the 4th International Conference on Human Language Technologies – The Baltic Perspective, Baltic HLT 2010*, Riga, Latvia, October 7-8, 2010.

XAlign(Alignement multilingue) LORIA (2006). http://led.loria.fr/outils/ALIGN/align.html.

MWEs: Support/light Verb Constructions vs Fixed Expressions in Modern Greek and French

Angeliki Fotopoulou
Institute for Language and
Speech Processing,
R.C. “Athena” www.ilsp.gr
afotop@ilsp.athena-innovation.gr

Voula Giouli
Institute for Language and
Speech Processing,
R.C. “Athena” www.ilsp.gr
voula@ilsp.athena-innovation.gr

Abstract

The paper reports on a study aimed at defining the limits between fixed expressions and Support Verb Constructions. To this end, a set of formal criteria that are applicable for the efficient classification of verbal MWEs were checked consistently against data in Greek and French. Ultimately, the delineation of the two types of constructions and their intermediate class is crucial not only for linguistic and lexicographic purposes, but also for Natural Language Processing tasks.

1 Introduction

Generally, verbal multi-word expressions (MWEs) fall into two general classes, namely idiomatic expressions and Support Verb Constructions (SVCs). Idioms or frozen/fixed expressions are defined as having a non-compositional meaning that cannot be deduced from the meaning of their parts (Bobrow & Bell, 1973; Chomsky, 1980; Fraser, 1970; Swinney & Cutler, 1979; M. Gross, 1982, 1988; Van der Linden, 1992). On the contrary, SVCs (also referred to in the literature as Light Verb Constructions, LVCs) consist of a support verb and a *predicative noun*. In between those two categories, however, a number of MWEs are proved to exhibit properties that are characteristic of both classes.

In this paper, we will try to discern the limits between fixed expressions and SVCs focusing on MWEs which fall in-between, comprising, thus, a grey-zone. The comparative study of MWEs in two languages, namely, Greek (EL) and French (FR), has a two-fold purpose: (a) to test the validity of the criteria set on the basis of cross-lingual similarities; and (b) to draw conclusions that might be useful for a range of NLP applications, either in single- or multi-lingual settings.

The paper is outlined as follows: section 2 briefly outlines the observations that motivated the present study; initial definitions and the criteria for disambiguation are presented in section 3. EL and FR data and the tests applied on them are presented in section 4; conclusions are discussed in section 5, whereas final conclusion and future research are outlined in section 6.

2 Motivation and Scope

The present study was triggered by two observations. On the one hand, the distinction between SVCs and fixed expressions is not always easy or straightforward and the limits between the two are often fuzzy. One could even maintain that there is a visible *scalar passage* between the two types of structures. In other words, a number of expressions seem to bear properties normally inherent to SVCs despite of their primarily being classified as fixed expressions and vice-versa. This is better illustrated in sentence (1) below taken from M. Gross (1981):

(1) il y a de l’eau dans le gaz
lit. it has of the water in the gaz
things are not running smoothly

The constituents of the expression in (1) cannot be modified if the meaning of the sentence is to be preserved; nevertheless, they allow for a paraphrase with the operator verb *mettre* (=put), that is a structure typical of SVCs:

(2) la venue de Max a mis de l'eau dans le gaz
*lit.*the arrival of Max has put of the water in the gaz
Max's arrival has caused irritation

On the contrary, a paraphrase with the verb *être* (=to be) is ungrammatical:

(3) *de l'eau est dans le gaz
*lit.*of the water is in the gaz

One step further, evidence from FR and EL data shows that the phenomenon is not language-specific. So the current study might prove to be useful for the analysis and representation of a phenomenon which is quite important for Greek - and not at all negligible for French and other languages alike. Based on these observations above, further *criteria* for the distinction between the two classes need to be elaborated, complementing, thus, the existing ones. These criteria are formal, depicting, thus, syntactic properties of MWEs.

3 Fixed Expressions and SVCs

Although there is nothing exceptional in their syntactic behavior or in their lexical content, fixed expressions are considered to be exceptions to the normal rules of the language. In an attempt to identify fixed expressions, linguistic tests are employed. The primary one is intuitive yet sufficiently operational: the meaning of fixed expressions is *non-compositional*, i.e. it cannot be computed from the meaning of its parts. The second test checks whether the constituents of the expression can be modified. At least two elements of a fixed expression do not allow for modification *(non-modifiability)*; usually, one of the two is the verb. These criteria constitute a mechanism for distinguishing between a *fixed* expression and a *free* one.

SVCs, on the other hand, are defined as expressions comprising a *support verb (Vsup)* (έχω/avoir (=have), παίρνω/prendre (=take), χάνω/perdre (=miss) in EL and FR respectively) or the *operator verb* δίνω/donner (=give) and a *predicative noun (Npred*) in *object*, *subject* or *PP complement* position. SVCs are defined as "semi-phrasal expressions formed by two lexical items l_1 and l_2 in which l_2 is taken in an arbitrary way to express a given sense and/or syntactic role in function of l_1's choice" (Alonso Ramos, 2004). In this respect, their semantics is more or less transparent and their meaning is *semi-compositional*.

A systematic treatment of SVCs can be found in Gross (1981) and Danlos (1986), Vives (1993) in French. According to these studies, although highly idiosyncratic, SVCs exhibit regularities. For example, a nominal predicate forming a SVC with the Vsup *avoir* (=have) can also form aspectual variants when combined with other support verbs, i.e., prendre (=take), perdre (=loose), etc. or operator and causative verbs. An example of this type of expressions is shown in Table 1 below.

EL	FR	EN
έχω κουράγιο *echo kurajio* to-have$_{\text{1.SING}}$ courage	avoir de courage	*(=have courage)*
παίρνω κουράγιο *perno kurajio* to-take$_{\text{1.SING}}$ courage	prendre du courage	*(=take courage)*
χάνω το κουράγιο μου chano to kurajio mu to-loose$_{\text{1.SING}}$ the courage my$_{\text{poss.1.SING.GE}}$ to-loose$_{\text{1.SING}}$ my courage$_{\text{.SING.ACC}}$	perdre son courage	*(=loose my courage)*
δίνω κουράγιο σε δino kurajio se give$_{\text{1.SING}}$ courage to	donner du courage à	*(=give courage to)*

Table 1. Examples of SVCs

Finally, the SVC is semantically equivalent to a full verb and can, thus, be paraphrased as an elementary sentence; in this context, the *Vsup* licenses *nominalisation*, a term which, in this context, refers to the relation between two sentences, one of which is a verbal construction (as shown in (4) below) and the second one is a SVC (depicted in (4a):

(4) Marie *respecte* son père
Marie respects her father

(4a) Marie *a du respect* pour son père
(Marie has respect for her father)

The relation that holds between the two elementary sentences[i] (4) and (4a) is of the form:

N0 V N1 = N0 Vsup Det Npred Prep N2[ii]

The implication is two-fold: (a) in terms of semantics, the *Vsup does* contribute to the meaning of the expression, *maintaining*, thus, some sort of lexical meaning; and (b) even though there is a great deal of idiosyncrasy in the formation of SVCs, regularities with respect to the alternations are attested.

However, as we will show below (examples 8a- 8c) grey zones arise when these regularities –

alternations are attested in expressions initially classified as idiomatic and vice versa. In this context, the intuitive criterion of compositionality seems to be insufficient for the robust classification of MWEs; to this end, we test a set of formal criteria that complement the semantic criterion.

3.1 Criteria for disambiguation

In the following, we will present the *structural properties* of SVCs; these may be further employed as formal criteria for disambiguation when classification seems to be problematic. More precisely:

(i) an elementary sentence of the form *N0 V N1* or *N0 V Prep N1* may be classified as *SVC if it can be paraphrased* as a *Vsup sentence*, that is, one that comprises one of the most common *Vsup*, namely *έχω/avoir* (=have) and *είμαι/être* and a predicative noun (Npred):

N0 V N1 = N0 (έχω/avoir+είμαι/être) Npred

Our hypothesis can then be formulated as follows: a MWE (which is otherwise difficult to be classified due to its exhibiting also properties of fixed expressions) falls into the class SVC, given that it can be associated with a simple *Vsup* construction and a Prep besides the predicative Noun (N). If a typical SVC can replace the original sentence, then the expression in question is considered to be an *aspectual* or *lexical* variant of a SVC. Otherwise, if the *Vsup* sentence is ungrammatical, the construction is considered to be a proper fixed expression instead.

(ii) An elementary sentence of the form *N0 V N* (or N0 *V Prep N*) is classified as SVC if an equivalence relation is proved to hold between this sentence and a structure of the form *Npred de N0* for French and *Npred N0$_{GEN}$ for* Greek; the latter structures are the result of a nominalization transformation. According to this requirement, therefore, if the *so derived nominal group* is *acceptable,* then the candidate expression is classified as SVC otherwise, it is a fixed expression.

4 The Data

The afore-mentioned tests were applied on EL and FR datasets developed within the Lexicon-Grammar framework (Gross, 1982), (Fotopoulou, 1993). The dataset comprises c. 1020 MWEs in EL c. 3700 expressions in FR. Iterative checks were performed over the EL and FR data to accurately perform MWE classification. In this regard, a number of structures were observed as displaying the characteristics of both MWE classes. Disambiguation, of grey expressions was guided by the application of the formal tests proposed in section (3.1) above.

In the remaining, we will present the cases of disambiguating MWEs based on the criteria proposed, i.e., capitalizing on the properties of two types of MWEs: (a) those with the Vsup έχω/avoir (=have) and (b) verbal MWEs with είμαι/être (=to be) and their variants.

4.1 Classification: έχω/avoir (=to have) test

Standard (lexical) tests employed for the identification of SVCs show that a number of verbs function as support verbs. This property, however, can be blocked when these verbs are used with specific nouns. For example, the verb *χαίρω* (=to enjoy) is used in SVCs of the form:

(5) Ο Νίκος χαίρει σεβασμού
o Nikos cheri sevasmu
the$_{SING.Nom}$ Nikos$_{SING.Nom}$ enjoys$_{3.SG}$ respect$_{SING.Gen}$
Nikos is respected.

However, the expression *χαίρω άκρας υγείας* (=to be healthy) was identified as a fixed expression on the basis of the criteria defined above. The test employed was of the form *Vsup C = (έχω/avoir) C.* As a matter of fact, the noun *υγεία* (=health) along with its modifier *άκρα* (=extreme) form together a unique combination that allows for no modification; the formation of a *Vsup* construction with the verb έχω (=have) results in an unacceptable construction. The Vsup paraphrase depicted in (5b) is unacceptable:

(5a) ο Νίκος χαίρει άκρας υγείας
o Nikos cheri akras igias
the$_{SING.Nom}$ Nikos$_{SING.Nom}$ enjoys$_{3.SG}$ [extreme health]$_{SING.Gen}$
Nikos is healthy

(5b) *ο Νίκος έχει άκρα υγεία
o Nikos echi akra igia
*lit.*the$_{SG.Nom}$ Nikos$_{SG.Nom}$ has$_{3.SG}$ extreme health
Nikos has extreme health

The application of the first test, i.e. the substitution of the verb by *έχω* (=have) has proven the sentence to be fixed. Along the same lines, the operation implied by the second criterion above is applied, i.e., elimination of the

Vsup and formation of a nominal group; the resulting structure is also unacceptable:

(5c) * η άκρα υγεία του Νίκου
i akra igia tu Niku
the$_{SG.Nom}$ extreme health$_{SG.Nom}$ of-the$_{SG.Gen}$ Nikos$_{SG.Gen}$
Niko's extreme health

However, when the modifier *άκρα* (=extreme) is replaced with another synonym, e.g. *εξαιρετική* (=excellent), paraphrasing the sentence with the use of έχω (=have) is permitted:

(5d) ο Νίκος χαίρει εξαιρετικής υγείας
o Nikos cheri ekseretikis igias
the$_{SG.Nom}$ Nikos$_{SG.Nom}$ enjoys$_{3.SG}$ excellent health
Nikos enjoys excellent health

(5e) ο Νίκος έχει εξαιρετική υγεία
o Nikos echi ekseretiki igia
the$_{SG.Nom}$ Nikos$_{SG.Nom}$ has$_{3.SG}$ excellent health
Nikos has excellent health

Now, when the modifier is replaced, the formation of the nominal group is allowed:

(5f) η εξαιρετική υγεία του Νίκου
i ekseretiki igia tu Nikou
the$_{SG.Nom}$ [excellent health]$_{SG.Nom}$ of-the$_{SG.Gen}$ Nikos$_{SG.Gen}$
Nikos' good health

We therefore conclude that on grounds of formal criteria, the expression in sentence (5a) is a fixed one. On the contrary, based on the criteria proposed, the expression *τρέφω ελπίδες* (=nourish hopes) was proved to be a pseudo-fixed expression since properties of SVCs were identified:

(6) Ο Γιάννης *τρέφει* ελπίδες …
O Jianis trefi elpiδes
lit. the John feeds$_{3.SG}$ hopes for
John hopes…

(6a) Ο Γιάννης *έχει* ελπίδες …
O Jianis echi elpiδes
lit. the John has$_{3.SG}$ hopes
John hopes…

(6b) οι ελπίδες του Γιάννη
i elpiδes tu Jiani
*lit.*the hopes of-the$_{SG.Gen}$ John$_{SG.Gen}$
John's hopes

Evidence from French follows:

(7) *Max voue* une admiration à
*lit.*Max pays admiration for
Max admires

(7a) Max a de l' admiration pour
*lit.*Max has admiration for
Max admires

(7b) l'admiration de Max
*lit.*the admiration of Max
Max's admiration

Nevertheless, the expression *παίρνω σάρκα και οστά* (=become true) which includes the *Vsup παίρνω* (=take), doesn't meet the formal criteria. As a matter of fact, neither the formation of the nominal group is acceptable, nor can the verb *παίρνω* (=take) be substituted by έχω (=have):

(8) το έργο παίρνει σάρκα και οστά
to erjo perni sarka ke osta
the$_{SG.Nom}$ project$_{SG.Nom}$ takes$_{3.SG}$ flesh and bones
the project starts

(8a) *σάρκα και οστά του έργου
sarka ke osta tu ergu
flesh and bones of-the$_{SG.Gen}$ project$_{SG.Gen}$
flesh bones of the project

(8b) *το έργο έχει σάρκα και οστά
to ergo echi sarka ke osta
the project has flesh and bones

However, a semantically similar expression that includes the causative operator *δίνω* (=give) is also possible:

(8c) η κυβέρνηση δίνει σάρκα και οστά στο έργο
i kivernisi dini sarka ke osta sto ergo
*lit.*the government gives$_{3.SG}$ flesh and bones to the project
the government puts flesh on the project

The same is attested in French expressions with lexical variations. The application of the two tests on the French expression *donner corps* results in a grammatical sentence when paraphrased with a simple *Vsup* (9a) and an ungrammatical nominal group (9b):

(9) Je donne corps à un projet
*lit.*I give body to a project
I give flesh to a project

(9a) le projet prend corps
*lit.*the project takes body
the project comes into being

(9b) * le corps du projet
*lit.*the body of-the project

4.2 Classification: είμαι/être (=to be) test

A number of expressions in our dataset seem to be paraphrases of structures displaying the form *είμαι/être (=to be) Prep C1*. Their aspectual variants contain either a verb that is prototypically defined as denoting movement (Vmt) or movement causative verb (Vcmt). To better account for properties these expressions have, a reference to the properties of *Vmts* and *Vcmts* is in order. These are depicted in examples (10)-(12) below.

N0Vmt Prep N1

(10) ο Νίκος πήγε στην εξοχή
o Nikos pige stin eksochi
the Nikos went to-the countryside
Nikos went to the countryside

(10a) Nikos est allé à la campagne
Nikos went to the countryside

N0 Vcmt N1Prep N2

(11) η Μαρία έστειλε τον Νίκο στην εξοχή
i Maria estile ton Niko stin eksochi
lit. the Maria sent Nikos to-the countryside
Mar;ia sent Nikos to the countryside

(11a) Maria a envoyé Nikos à la campagne
Maria sent Nikos to the countryside

N1είμαι(=to be)PrepN2

(12) ο Νίκος είναι στην εξοχή
o Nikos ine stin eksochi
lit. the Nikos is at-the countryside
Nikos is at the countryside

(12a) Nikos est à la campagne
Nikos is at the countryside

The relation that holds between simple structures with *Vmt* predicates of the type (10), (10a) and *Vcmt* structures (11), (11a) involves a semantic alternation. Sentences with a *Vmt* or a *Vcmt* have a *dynamic aspect*, whereas sentences of the form *είμαι (=be) Prep* have a *static* one. The dynamic aspect of the former can also be ascribed, in case of (11), (11a), to the causative operator that introduces the subject of the sentence. The same phenomenon is also attested in MWEs of the form *be Prep X* and their *variants comprising* verbs that are prototypically *Vmt* or *Vcmt* predicates, as illustrated in examples (13)-(18a) below:

N0 Vcmt N1Prep C2

(13) η Μαρία φέρνει τον Νίκο σε δύσκολη θέση
i Maria ferni$_{3.SG}$ ton Niko se diskoli thesi
lit. the Maria brings Nikos to difficult position

(13a) Marie a mis Nikos dans une situation difficile
lit. Maria put Nikos in situation difficult
=Maria made Nikos feel uncomfortable

N1Vmt Prep C2

(14) ο Νίκος ήρθε σε δύσκολη θέση
o Nikos irthe se δiskoli thesi
lit. the Nikos came to difficult position
Nikos was uncomfortable

(14a) Nikos sort d'une situation difficile
Nikos fate of a difficult situation

N1 είμαι (=be) PrepC2

(15) Ο Νίκος είναι σε δύσκολη θέση
O Nikos ine se δiskoli thesi
lit. the Nikos is in difficult position

(15a) Nikos est dans une situation difficile
lit. Nikos is in a situation difficult
Nikos is in a difficult position

N0 Vcmt N1Prep C2

(16) η ένταση οδήγησε τις διαπραγματεύσεις σε αδιέξοδο
i entasi oδijise tis diapragmatefis se aδιeksoδo
the conflict led the negotiations to an impasse

N1Vmt Prep C2

(17) οι διαπραγματεύσεις κατέληξαν/έφτασαν σε αδιέξοδο
i δiaprajmatefsis kateliksan/eftasan se aδieksoδo
the negotiations reached/arrived at an impasse

N1είμαι Prep C2

(18) οι διαπραγματεύσεις είναι σε αδιέξοδο
i δiapragmatefsis ine se adieksoδo
lit. the negotiations are at impasse
the negotiations are at an impasse

(18a) les négociations sont dans l'impasse
the negotiations are at an impasse

The same regularity is also manifested in the following examples:

N0 Vcmt N1Prep C2

(19) Ο Νίκος έβγαλε τη Μαρία από τη μέση
O Nikos evjale ti Maria apo ti mesi
*lit.*The Nikos took the Maria from the middle
Nikos took Maria out of the way

N1Vmt Prep C2

(20) Η Μαρία βγήκε από τη μέση
i Maria vjike apo ti mesi
lit. the Maria was taken from the middle
Maria was taken out of the way

The expressions in (19) and (20) are MWEs bound with the alternation implied by their *Vmt* and *Vcmt* predicates, even though the *be Prep X* structure has a compositional meaning:

N0 be Prep X

(21) Η Μαρία είναι στην μέση
i Maria ine sti mesi
lit. the Maria is in-the middle

5 Discussion

Summing up the presentation of the different cases, we observe two main classes of MWEs: (i) fixed expressions, which allow for regular syntactic alternations, and which do not correspond to a simple SVC; and (ii) fixed expressions that follow the *be Prep X* paradigm, in that they have aspectual variants or causative operations, yet they lack the basic *be Prep X* form. In our view, an efficient computational treatment of MWEs within NLP applications requires a detailed analysis of each expression and the definition of their so-called *fixed zone* (zone fixe) - a term coined to denote a fixed set of words or tokens within an MWE that would be optionally amenable to *morphological* variation. In this sense, the support verb within a support verb construction (SVC) is not considered to be part of the SVC *fixed zone* (Laporte, 1998). However, variation within the SVC can also be lexical/aspectual. It is, therefore, crucial to redefine and delineate the notion of fixed zone and to define more elaborate linguistic features that apply to MWEs regardless their initial classification. This will have an impact on their treatment in NLP. For example, expressions that do not correspond to the SVC paradigm, but, nevertheless, allow for certain operations should be encoded as appropriate.

6 Conclusion

This study is aimed at delineating the boundaries between fixed expressions and *SVCs* constructions (or collocations) in two languages, namely French and Greek. A set of formal tests suitable for the classification of verbal MWEs have been checked. The focus is primarily on expressions that seem to exhibit features inherent to both fixed expressions and SVCs. Future work involves the application of these tests to ultimately all MWEs in the Lexicon-Grammar tables and beyond, aiming at the elaboration of an accurate and consistent classification thereof. The latter is particularly important in view of processing MWEs in a number of applications including alignment and paraphrasing, translation, etc.

References

Alonso Ramos, Margarita. 2004. Las construcciones con verbo de apoyo. Madrid: Visor.

Bobrow, S.A. & Bell. S.M. 1973. On Catching on to Idiomatic Expressions. Memory and Cognition 1 (3): 343-346.

Chomsky, N. 1980. Rules and Representations. New York: Columbia University Press.

Fraser, B. 1970. Idioms within a Tranformational Grammar. Foundations of language 6 (1): 22-42.

Fotopoulou, A., 1992. Dictionnaire électronique des phrases figées: traitement d'un cas particulier : phrases figées - phrases à Vsup. COMPLEX '92, Budapest, Hungary.

Gavrilidou, M. & Fotopoulou, A. A special class of complex predicates: Frozen predicates. Technical Report Eurotra -El, ILSP Eurotra -Fr, L.I.S.H.

Gross. M. 1975. *Méthodes en syntaxe*, Paris: Hermann.

Gross, M. 1981. Les bases empiriques de la notion du prédicat sémantique, Langages 63, Larousse.

Gross, M. 1988. Les limites de la phrase figée. *Langages 23 (90): 7-22*.

Laporte, É., Nakov, P., Ramish, C., and Villavisencio, A., (eds) 2010. *Proceedings of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, Beijing, China. ACL.

Mel'cuk, I. 1998. Collocations and lexical functions. In Cowie, A.P. (ed.). *Phraseology*. *Theory, Analysis, and Applications,* 23–53. Oxford: Oxford University Press.

Nunberg, G., Sag, I., and Wasow, T. 1994. *Idioms*. *Language*, 70:491–538.

Ramos, M. A. 1999. Étude sémantico-syntaxique des constructions à verbe support. PhD Thesis, University of Montreal.

Ruwet, N., 1983. *Du bon usage des expressions idiomatiques dans l'argumentation en syntaxe générative.* Revue québécoise de linguistique 13 :1, Presses de l'Université du Québec à Montréal, Montréal.

Swinney,D. & Cutler,A. 1979. The Access and Processing of Idiomatic Expressions. *Verbal Learning & Verbal Behavior* 18(5):523-534.

Van der Linden, E.J. 1992. Incremental Processing and the Hierarchical Lexicon. *Computational Linguistics* 18 (2): 219-238.

Wehrli, E., Seretan, V., and Nerima, L. 2010. Sentence analysis and collocation identification. In [Laporte et al., 2010], pages 27–35

Vives, R. 1993. La prédication nominale et l'analyse par verbes supports. *L'Information grammaticale,* 59(1), 8-15.

Vives, R. & Gross, G. 1986. Les constructions nominales et l'élaboration d'un lexique-grammaire, Langue Française 69, Larousse, Paris

[i] The term *elementary sentence* is defined as the linguistic unit of meaning in accordance with Z. Harris' Transformational Grammar and the Lexicon-Grammar framework (Gross 1975)

[ii] The notation employed here is the one adopted by Lexicon-Grammar (Gross 1975).

Challenges on the Automatic Translation of Collocations

Ângela Costa
L2F/INESC-ID
CLUNL
angela@l2f.inesc-id.pt

Teresa Lino
CLUNL
tlino@mail.telepac.pt

Luísa Coheur
L2F/INESC-ID
IST
luisa.coheur@l2f.inesc-id.pt

Abstract

As a linguistic phenomenon, collocations have been the subject of numerous researches both in the fields of theoretical and descriptive linguistics, and, more recently, in automatic Natural Language Processing. In the area of Machine there is still improvements to be done, as major translation engines do not handle collocations in the appropriate way and end up producing literal unsatisfactory translations. Having as a starting point our previous work on machine translation error analysis (Costa et al., 2015), in this article we present a corpus annotated with collocation errors and their classification. To our believe, to have a clear understanding of the difficulties that the collocations represent to the Machine Translations engines, it is necessary a detailed linguistic analysis of their errors.

1 Introduction

According to (MELćUK, 1998), collocations are particularly relevant in the context of lexical combinatory as they are "the absolute majority of phrasemes and represent the main challenge for any theory of phraseology". (Tutin Agnés, 2002) defines them as "a privileged lexical co-occurrence of two (or more) linguistic elements that establish a syntactic relationship between them". (Hausmann, 1989), (Hausmann, 1985) and (Hausmann, 1984) observed that the status of the constituents are not similar, registering between them an hypotactic relationship. Hausmann calls "base" to the word that determines the choice of the co-occurring element and "collocate" the determined constituent.

The relationship between base and collocate is, in most cases, unpredictable, and does not demonstrate a particularly clear semantic motivation that can explain it. This idiosyncratic character and the fact that they cannot yet be considered lexicalized expressions, standing between lexicon and grammar, makes them very complex structures, from the production point of view. In fact, (Cruse, 2004) considers them "idioms of encoding", as they do not particularly cause problems from the decoding perspective, being relatively transparent constructions and syntactically regular. The problem lies on producing them, since the relationship between the base and collocate is, in most cases, arbitrary. Considering the translation task, we can imagine the number of problems that can occur, as a word-by-word translation may not always be the best choice. For instance, *break a record* cannot literally be translated into French *casser un record*, but as *battre un record* (lit. to beat a record).

In this article we briefly describe the role of collocations on machine translation, then we describe the corpus and error typology used in our study, finally we present the error analyses and the conclusions.

2 Collocations in Machine Translation

Collocations have been the subject of numerous researches both in the fields of theoretical and descriptive linguistics, and, more recently, in Natural Language Processing (NLP), as they can be useful for many language processing tasks, like parsing, word sense disambiguation, text generation and machine translation.

Although there are several methods for the extraction of collocations from corpora and evaluation of extraction results, the area of post-processing of this structures and their application to various branches of NLP is still at the beginning, especially in the area of machine translation (Seretan and Wehrli, 2007). Because of their semantic irregularities, collocations cannot always be translated word-by-word, creating a problem for automatic translation. In this example of a

Google translation the collocation *high wind* was literally translated to *vento alto* (lit. tall wind) instead of *vento forte*. On the other side, sometimes a literal translation may be correct, *make the bed* was translated to *fazer a cama* which is correct. Just as for a student learning a foreign language, also for an MT system is not always easy to know when the correct option is a word-by-word translation.

Error analysis of collocations in machine translation is still lacking. For instance two of the most used error taxonomies by (Bojar, 2011) and (Vilar et al., 2006) do not consider collocational errors on their classification. As previously mentioned, collocations have at least two elements, so the errors may concern any of the elements of the collocation (base, collocate) or the collocation as a whole. Finding the error within is compositional parts can help improve the translation of these structures.

3 Error Analysis

3.1 Corpus

Having as a starting point our previous work on machine translation errors (Costa et al., 2015), the error analysis of collocations was carried out on a corpus generated by four different systems: Google Translate[1] (Statistical), Systran[2] (Hybrid Machine Translation) and two in-house Machine Translation systems trained using Moses[3], and the two popular models: the phrase-based model (Koehn et al., 2007) (PSMT) and the hierarchical phrase-based model (Chiang, 2007) (HSMT), in three scenarios representing different challenges in the translation from English to European Portuguese:

- 250 sentences taken from TED talks[4];
- 250 sentences taken from the bilingual Portuguese national airline company: TAP magazine “UP”[5];
- 250 questions taken from a corpus made available by (Li and Roth, 2002), from the TREC collection (Li and Roth, 2002; Costa et al., 2012).

The Ted talks, in the original text in English had 3.346 tokens, the TAP and the corpus of Questions had 3.346 and 1.856, respectively. We were able to find a total of 172 collocations: 41 were found on the TED corpus, 84 on the TAP magazines and 47 on the Questions corpus. As previously mentioned, the three datasets were translated by four translation engines, so in total we have evaluated 164 collocations on the TED corpus, 336 on the TAP corpus and 188 on the Questions corpus.

3.2 Error types

To assess the errors that we have found, we used the location dimension of (Wanner et al., 2011) taxonomy to evaluate students errors when producing collocations. The first two categories show errors that were found on one of the two elements of the collocation (cf. (1) wrong collocate use and (2) wrong base use) and the third type problems that affected the collocation as a whole (cf. (3)).

1. **wrong collocate**: *cores preliminares*, lit. “preliminary colors” (instead of *cores primárias*, “primary colors”), *cabelo cinzento*, lit. “gray hair” (instead of *cabelo grisalho*, “gray hair”), *terra nativa*, lit. “native land” (instead of *terra natal*, “native land”)

2. **wrong base**: *perspectiva obtusa*, lit. “obtuse perspective” (instead of *ângulo obtuso*, “obtuse angle”), *começar uma faixa*, lit. “start a strip” (instead of *começar uma banda*, “start a band”), *meta cardíaca*, lit. “heart goal” (instead of *ritmo cardíaco*, “heart rate”), *flopped miseravelmente*, lit. “flopped miserably” (instead of *falhar miseravelmente*, “failed miserably”)

3. **wrong collocation**: *pagamento de separação*, lit. “payment of separation” (instead of *indemnização*, “compensation”), *ter ceia*, lit. “have supper” (instead of *jantar*, “have diner”)

   The errors found on a collocation can be rooted in the lexicon or in the grammar. A lexicon error concerning the base or the collocate consists in the incorrect translation of one of the two elements or both. This error can be caused by a literal translation from English that does not work in the context of the collocation, a near-synonym or even the non-translation of an element (see examples (1)

[1] http://translate.google.com
[2] http://www.systranet.com/translate
[3] http://www.statmt.org/moses
[4] http://www.ted.com/
[5] http://upmagazine-tap.com/

and (2)). When the error concerns the whole collocation, we found that new expressions with the structure of a collocation were created, meanwhile a single word should have been used (see examples (3)).

Grammatical errors can also affect the collocation as a whole or all of its parts (base and collocate). We were able to find four types: erroneous absence or presence of determiner, wrong number use, wrong order of the words and wrong government; cf:

4. **determiner**: *pedir a ajuda*, lit. "ask the help" (instead of *pedir ajuda*, "ask for help").

5. **number**: *mudar os canais*, lit. "change the channels" (instead of *mudar o canal*, "change the channel").

6. **reordering**: *chá de conjunto*, lit. "tea of set" (instead of *conjunto de chá*, "set of tea").

7. **government**: *sede para conhecimento*, lit. "thirst for knowledge" (instead of *sede de conhecimento*, "thirst for knowledge"), *carreira solo*, lit. "career solo" (instead of *carreira a solo*, "solo career").

4 Results

Figure 1 shows the number of errors present on each translation engine per error type. The correct translations are not represented on the graphic, but they were the majority of the cases, as Google, HSMT, PSMT and Systran produced 144, 114, 111 and 92 correct translation, respectively. From Figure 1, we can observe that:

- choosing the correct base of the collocation is not as problematic as deciding on the collocate, as this is the most common error for all engines;

- between 14% and 19% of the errors affect the collocation as a whole;

- determinant, number, reordering and government errors are not so common.

5 Conclusions

From this study we could observe that only between 14% and 19% of the errors affect the collocation as a whole. Determinant, number, reordering and government errors are not so common, as

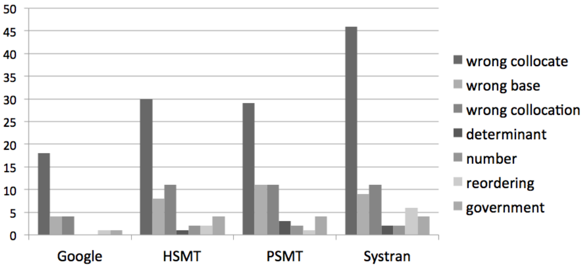


Figure 1: Number of collocation errors per system.

there is a relativity high congruence between English and Portuguese, although this may not be valid for other languages.

On all four MT systems the majority of the errors occur when choosing the collocate. This was also observed on foreign language learners on the already mentioned study by (Wanner et al., 2011). The source of the errors are literal translations of the collocate ("grey" - cinzento), use of a wrong synonym ("angle" - *perspectiva*) or untranslations (e.g. "flopped").

Although our analysed corpus is still very small, we think that it is a good contribution to have a clear understanding of the difficulties that the collocations represent to Machine Translations engines. Only after a detailed linguistic analysis of the errors, we can implement solutions, like finding and automatically correcting collocations.

Acknowledgments

This work is supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project UID/CEC/50021/2013. Ângela Costa is supported by a PhD fellowship from Fundação para a Ciência e a Tecnologia (SFRH/BD/85737/2012).

References

O. Bojar. 2011. Analysing Error Types in English-Czech Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, pages 63–76.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228, June.

Ângela Costa, Tiago Luís, Joana Ribeiro, Ana Cristina Mendes, and Luísa Coheur. 2012. An English-Portuguese parallel corpus of questions: translation guidelines and application in SMT. In *Proceedings of the Eight International Conference on Language*

Resources and Evaluation (LREC'12), pages 2172–2176, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Ângela Costa, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161.

D. A. Cruse. 2004. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press.

Franz Josef Hausmann. 1984. Wortschatzlernen ist kollokationslernen. zum lehren und lernen franzsischer wortverbindungen. *Praxis des neusprachlichen Unterrichts 31*, pages 395–406.

Franz J. Hausmann. 1985. Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In Henning Bergenholtz and Joachim Mugdan, editors, *Lexikographie und Grammatik: Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.-30.6.1984*, Lexikographica, pages 118–129. Max Niemeyer, Tübingen.

Franz Josef Hausmann. 1989. Le dictionnaire de collocations (artikel 95). In Wiegand H.E. Zgusta L. Hausmann F.J., Reichmann O., editor, *Wörterböcher - Dictionaries - Dictionnaries. Ein internationales Handbuch zur Lexicographie. Erster teilband.* Walter de Gruyter, Berlin/New York.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

X. Li and D. Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, pages 1–7. ACL.

Igor MELćUK. 1998. Collocations and lexical functions. *2001 [1998]*, pages 23–54.

Violeta Seretan and Eric Wehrli. 2007. Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, pages 401–410.

Grossmann Francis Tutin Agnés. 2002. Collocations régulléres et irréguliéres : esquisse de typologie du phénoméne collocatif. *Revue française de linguistique appliquée*, VII:7–25.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, pages 697–702, Genoa, Italy, May.

Leo Wanner, M Alonso Ramos, Orsolya Vincze, Rogelio Nazar, Gabriela Ferraro, Estela Mosqueira, and Sabela Prieto. 2011. Annotation of collocations in a learner corpus for building a learning environment. *Sylviane Granger/Gaëtanelle Gilquin/Fanny Meunier (edd), Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead. Corpora and Language in Use. Proceedings*, 1.

Chunking-based Detection of English Nominal Compounds

Gábor Csernyi
University of Debrecen
PO box 73.
Debrecen, H-4010, Hungary
csernyi.gabor@gmail.com

Abstract

This paper presents a supervised learning approach to detect English nominal compounds (NCs). The method, which is based on *chunking*, originates from identifying full noun phrases (as chunks) in POS-tagged texts, and shows that even basic syntactic information (in the form of POS tags) can be exploited to detect this type of *multiword expressions* (MWEs) with considerable results, even compared to dictionary-based and hybrid (the former way combined with machine learning) methods. The results of the experiments presented here also show how the size and the *density* of the training set (in terms of the frequency of the target expressions) as well as of the test set influence the efficiency of the algorithm(s).

1 Introduction

Nominal compounds (NCs), a (sub)type of multiword expressions (MWEs) have been widely explored recently. These special linguistic phenomena are lexical items, and as it has been shown, they are quite frequent in any language. They reflect idiosyncratic features: the words making up a nominal compound – while each having their own meaning – together form a single expression that functions as a noun (Sag et al., 2002; Nagy T. et al., 2011). No matter these expressions are compositional in meaning or not, they are to be treated as a *single semantic unit*. The importance of identifying them in running texts is therefore further underlined by fields as information retrieval/extraction as well as machine translation, where it is crucial to detect them each as a single semantic whole. However, it should also be noted that not every nominal combination or co-occurrence functions as a nominal compound (like *fat cat*, where *fat* can be an adjectival modifier of *cat* in one context, while in other cases the whole expression can also mean *well-paid executive*), and it is generally the context that might help us decide if the compound candidate is a real compound (Nagy T. and Vincze, 2013). Furthermore, since these expressions are quite productive, they do not constitute a fixed subset of the language; new terms of this type might appear in the language anytime (Nagy T. and Vincze, 2013).

From the nature of compounds listed above, taking the natural language processing perspective of automatic detection, it also follows that our focus of interest concerns those cases where the parts of compounds are delimited by a space; hyphenated compounds or those in which the parts are written together do not necessarily pose problems concerning identifying them as a unit.

This paper reports on an alternative machine learning approach to identify nominal compounds with the help of chunking. Section 2 provides a very brief overview of related works and their results, while Section 3 presents the experiment including the corpora and the exact methods in a detailed way, reflecting on the relevant features of the former, and the efficiency of the latter, as well. In Section 4 concluding remarks follow.

2 Related Work

The notion of nominal compounds (and multiword expressions in general) has been getting more and more focus from the natural language processing perspective recently. Alignment-based NC detection approaches using parallel texts for English and Portuguese have been reported by Caseli et al. (2009). Bonin et al. (2010) rely on contrastive filtering to extract NCs from various texts including Wikipedia articles. Wikipedia as a resource is also exploited by Nagy T. and Vincze (2013) in that they automatically compile NC dictionaries from Wikipedia articles. They also show that these

NC dictionaries, combined with machine learning methods (using conditional random fields classifiers) can improve the performance of such systems[1].

Another machine learning tool available for NC extraction is the `mwetoolkit`, especially created to extract MWEs (Ramisch et al., 2010a). Relying on this toolkit Ramisch et al. (2010b) presents case studies to indentify nominal compounds in generic-purpose as well as in domain-specific texts.

This paper takes an alternative (though not a novel) approach and shows how an existing tool, which by design operates on word category level, can be applied to NC detection.

3 The Experiments

The experiments described in this paper utilize the algorithms originally used for noun phrase chunking (i.e. detecting and extracting full noun phrases using the output of a POS tagger) detailed by Bird et al. (2009), also in connection with the CoNLL-2000 Shared Task. Among the experiments presented here, there are two rule based methods that rely on extraction via regular expressions (to serve as basis of comparison), while the machine learning approaches involve consecutive chunkers with various settings with regard to what features to consider in the classification process(es) carried out by the NLTK Megam maximum entropy model[2].

3.1 The corpora

The models to be demonstrated make use of two corpora. The training set is the Wiki50 corpus (Vincze et al., 2011) comprising 50 English Wikipedia articles, each with at least 1000 words. The size of this corpus is altogether 114570 tokens, there are 2929 noun compound occurrences (2405 unique phrases) in it, and out of the 4350 sentences (that is the full size of the set) 1931 contain at least one NC .

The test set is the BNC dataset of 1000 sentences compiled by Nicholson and Baldwin (2008), which is a selection of 1000 sentences from the British National Corpus, containing annotations for two-word NCs primarily[3]. The number of NCs – which are actually nominalizations – is 358, subcategorized according to whether they (as nominalizations) have subject or object interpretations (however, these subcategories are disregarded in the present study). In this corpus the number of sentences that contain a minimum of one NC is 267. These statistics related to the two corpora are shown in Table 1.

	Wiki50	BNC dataset
size (sentences)	4350	1000
tokens	114570	~19100
NCs	2929	358
sentences with NCs	1931 (44.39%)	267 (26.7%)

Table 1: Corpus statistics.

In their original state both sets are (manually) tagged for multiword expressions only, and contain no part-of-speech annotations that the experiments would use. The formats of the sets are also different. Wiki50 uses annotations in separate files, one for each of the 50 texts, and these annotation files contain pointers with regard to the positions of the multiword items in the original documents. Concerning the BNC dataset, it is in XML format. Although expressions other than NCs are also marked in both sets, they are disregarded with respect to the current experiments (including multiword named entities, as well).

3.2 The methods

As for chunking processes, there is a need for POS tags, and also, corpora are to be stored in IOB format (in which one line takes a token, its POS tag, and an I, an O, or a B mark depending on whether the token in question is inside, outside, or is the beginning of a nominal compound) following Bird et al. (2009). Consequently, both sets are POS-tagged first with the standard POS tagger of the Natural Language Toolkit (NLTK), and are also converted into the desired IOB format[4].

There are four settings to which the training and testing algorithms are applied. These settings differ in terms of the size of the two (training and

[1] They demonstrate their findings evaluated on the Wiki50 corpus and the BNC dataset of 1000 sentences (which are also the target corpora of the experiments detailed in this paper).

[2] Available from `http://goo.gl/1UPtBX`.

[3] This is a fact that might have a negative effect on the rersults of models trained on longer expressions.

[4] Conversion tools available from `https://goo.gl/LIKSAL` for Wiki50 and from `https://goo.gl/1KWBVu` for the BNC dataset.

test) corpora and therefore the density of the target expressions in them:

- **setting 1** includes only filtered corpora, i.e. only those sentences are taken from both the test and the training set which contain at least one nominal compound;
- **setting 2** takes the full training set (that is all the sentences from the given corpus) and the filtered test set;
- **setting 3** consists of a filtered training set and the full test set;
- **setting 4** is the scenario in which the full version of both sets are taken.

In the full training set only 5.6% of the tokens are NC tokens, in the filtered one this measure increases to 10.8%. Regarding the test set, 3.5% of all the tokens belong to NCs, while in the filtered version this frequency value reaches 9.8%. Therefore, comparing the two corpora in terms of their size and the distribution of NCs within them reflects similar ratios. However, these values also show that even if around 44.4% of the training set sentences and 26.7% of the test sentences (concerning the full sets) contain at least one NC, their number is relatively low as compared to the total number of tokens.

Each of the algorithms discussed above are exploited with settings 1-4 mentioned above. As for the consecutive chunker, which carries out the classification process with the maximum entropy model, four distinct feature sets ranging from the simple candidate form to its wider contexts are taken care of in the four different configurations deltailed above. The first chunker (*Consecutive 1*) focuses on the POS tag of the current and the previous token only. The second (*Consecutive 2*) also pays attention to the token as a part of a compound noun itself. In the third configuration (*Consecutive 3*) the previous feature sets are extended by the next token and its POS tag, and pair-combinations of the current and the two noncurrent POS tags (previous+current, and current+next). In this latter feature set the distance of the current token from the last determiner is also taken into account. This feature is disregarded in the last configuration (*Consecutive 4*).

	IOB	**P**	**R**	**F**
<NN><NN>+	93.3	65.4	41.7	50.9
<[JN].><NN>+	89.2	33.5	37.3	35.3

Table 2: RegExp search patterns for NCs within the filtered BNC dataset.

	IOB	**P**	**R**	**F**
<NN><NN>	96.6	43.8	40.9	42.3
<[JN].><NN>	92.3	16.6	36.1	22.7

Table 3: RegExp search patterns for NCs within the full BNC dataset.

3.3 Results

Using simple regular expressions to extract NC candidates from the filtered BNC dataset reaches quite low f-scores, as Table 2 show. Although this corpus is primarily tagged for two-word compounds, in some instances annotated compounds of more than two words can also occur. Another factor to be considered in connection with the performance of the regular expression parsers is that the standard POS tagger might mark the initial element(s) of a compound as adjective (that is, after the POS tag NN, the second most frequent case), or sometimes participle, by mistake. In addition, it should also be noted that regular expressions work in linear order (taking larger chunks, or excluding one pattern if another fits), which can be an explanation to why the recall values are so low. Comparing the efficiency of two regular expressions, one accepting noun chains, and the other also accepting an adjective as the initial component of the candidate NC, the former covers more examples and with higher precision, nevertheless, the whole coverage is rather below expectations. The situation is even less satisfactory with the full BNC dataset, however, it is still the regular expression ignoring adjectives as part of NCs that yields a higher f-score, as Table 3 suggests.[5].

The chunkers *Consecutive 1* and *Consecutive 2* focusing on minimal (or no) context do not seem to cope with the problem very well either, in any of the settings. Applying those chunkers that consider larger contexts of the candidate NCs, however, appear to be much more efficient, even compared to the regular expression searches (as can be

[5] The difference between precision and IOB-accuracy is that the former is related to chunks (just like recall) while the latter is to tokens.

	IOB	P	R	F
Consecutive 1	89.8	4.1	3.9	4.0
Consecutive 2	89.8	6.8	7.0	6.9
Consecutive 3	94.9	60.6	67.5	63.8
Consecutive 4	95.0	60.7	67.8	64.0

Table 4: Results of consecutive chunkers in setting 1 (Wiki50-filtered, BNC_dataset-filtered).

	IOB	P	R	F
Consecutive 1	89.8	4.1	3.9	4.0
Consecutive 2	89.8	4.1	3.9	4.0
Consecutive 3	95.1	61.6	61.1	61.3
Consecutive 4	94.9	60.0	59.7	59.8

Table 5: Results of consecutive chunkers in setting 2 (Wiki50-full, BNC_dataset-filtered).

	IOB	P	R	F
Consecutive 1	95.7	3.0	3.9	3.4
Consecutive 2	95.5	4.7	7.0	5.7
Consecutive 3	96.6	40.3	67.5	50.5
Consecutive 4	96.7	41.4	67.8	51.4

Table 6: Results of consecutive chunkers in setting 3 (Wiki50-filtered, BNC_dataset-full).

	IOB	P	R	F
Consecutive 1	95.7	3.0	3.9	3.4
Consecutive 2	95.6	3.0	3.9	5.7
Consecutive 3	97.2	44.4	61.1	51.4
Consecutive 4	97.1	43.0	59.7	50.0

Table 7: Results of consecutive chunkers in setting 4 (Wiki50-full, BNC_dataset-full).

seen in the last two rows Table 4-5 and Table 6-7). Although the highest recall is still lower than that of the noun chain RegExp extractors, the precision is far better, and these chunkers, *Consecutive 3* and *Consecutive 4* reach higher f-scores, as a result.

Another observation in connection with these last two chunkers is that using the full set for training, the method that does not pay attention to the distance between the first token of the candidate NC and the last determiner preceding it (*Consecutive 4*) performs better both in terms of precision and recall, no matter the test set is the full or the filtered one. In contrast, training the models on the filtered set, the chunker focusing on this distance feature (*Consecutive 3*) as well yields slightly better results.

4 Conclusion

Concerning the size of the corpora in terms of the frequency of NCs in them, as could be predicted, the denser the training and the test set in NCs, the more efficient these chunkers might turn out to be. There is a considerably high difference regarding the performance of these machine learning approaches when they are applied to the filtered and to the full training or test sets. Regardless of the fact that filtering comes with even smaller sets, these relatively small corpora can still function as useful language resources to train and test the algorithms and similar ones to run experiments extracting/detecting NCs. An extension to the test corpus could be to annotate NCs of more than two words, which could give a more precise measure of the performance of these chunkers trained on such examples, as well.

From among the different approaches, consecutive chunkers reach the highest overall scores. As it can be seen, the more features these classifiers take, the better their coverage of detecting NCs are. However, apart from the limited set of features considered here, further contextual and semantic characteristics could also be exploited (e.g. dependencies), which might enhance the performance of the models.

It must be added, however, that the machine learning approaches described here all rely on a simple pointwise maximum entropy model, thus a maximum entropy Markov model (MEMM), or a conditional random field (CRF) model – both of which are quite powerful when working with larger feature sets – would probably reach higher scores. It would also be desirable to try other standard POS taggers, like the Stanford POS tagger, since the output of these tools definitely influence how the models are trained and how they perform then. Testing the trained models on larger datasets, like on the Wall Street Journal component of Penn Treebank, could also provide a represenative baseline, another possible research direction in terms of benchmarking, which might help make these parsers more powerful in terms of indentifying NCs.

References

Steven Bird, Ewan Klein and Edward Loper. 2009. *Natural Language Processing with Python: An-*

alyzing Text with the Natural Language Toolkit. O'Reilly, Beijing.

Francesca Bonin, Felice Dell'Orletta, Giulia Venturi and Simonetta Montemagni. 2010. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*. 77-80.

Helena de Medeiros Caseli, Aline Villavicencio, Andr Machado and Maria Jos Finatto. 2009. Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications.* ACL, Singapore. 1-8.

István Nagy T., Gábor Berend and Veronika Vincze. 2011. Noun Compound and Named Entity Recognition and their Usability in Keyphrase Extraction. In *Proceedings of Recent Advances in Natural Lanugage Processing 2011* (RANLP 2011). Hissar, Bulgaria. 162-169.

István Nagy T. and Veronika Vincze. 2013. English Noun Compound Detection With Wikipedia-Based Methods. In Václav Matousek, Pavel Mautner, Tomás Pavelka (eds.), *Proceedings of the 16th Inter-national Conference on Text, Speech and Dialogue* (TSD 2013), *Lecture Notes in Computer Science* . Springer, Berlin / Heidelberg. 225-232.

Jeremy Nicholson and Timothy Baldwin. 2008. Interpreting Compound Nominalisations. In *Proceedings of LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions* (MWE 2008). Marrakech, Morocco. 43-45.

Carlos Ramisch, Aline Villavicencio and Christian Boitet. 2010a. mwetoolkit: a framework for multiword expression identification. In In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC 2010). Valetta, Malta, May. European Language Resources Association.

Carlos Ramisch, Aline Villavicencio and Christian Boitet. 2010b. Web-based and combined language models: a case study on noun compound identification. In In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias (eds.), *Coling 2010: Poster Volume*. 1041-1049.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING 2002). 1-15.

Veronika Vincze, István Nagy T. and Gábor Berend. 2011. Multiword Expressions and Named Entities in the Wiki50 Corpus. In *Proceedings of Recent Advances in Natural Lanugage Processing 2011* (RANLP 2011). Hissar, Bulgaria. 289-295.

Integrating Multi Word Expressions in Statistical Machine Translation

Zied Elloumi
LIG - GETALP
Univ. Grenoble Alpes
zied.elloumi@imag.fr

Laurent Besacier
LIG - GETALP
Univ. Grenoble Alpes
Laurent.Besacier@imag.Fr

Olivier Kraif
LIDILEM
Univ. Grenoble Alpes
olivier.kraif@u-grenoble3.fr

Abstract

Multiword Expressions (MWE) are complex phenomena which represent a real challenge for computational linguistics, and Statistical Machine Translation (SMT) systems in particular. In this paper, we try to improve translation of Multiword Expressions in SMT. Preliminary results, regarding the BLEU score, are encouraging.

1 Introduction

Multiword expressions (MWEs) processing is a highly important research field in computational linguistics. These expressions are formed by combining two or more words, and they cover a large number of phraseological phenomena that ranges from collocations to phrasal verbs (PV), idioms, etc. However, these expressions are a challenging issue for machine statistical translation systems, since their meanings cannot be easily predicted from the words they contain (non compositionality).

The examples below show the type of errors made by MT systems when facing MWEs:

- Example 1 : ***Phrasal verbs***

Source:
« He picked up a remote and turned the stereo on. »
Google-translate:
« Il ramassa une télécommande et tourna la stéréo sur. »

→ The correct translation of phrasal verbs is *« alluma »*

- Example 2 : ***Idioms***

Source:
« my job was a piece of cake. »
Google-translate:
« mon travail était un morceau de gâteau. »

→ the correct translation should be : *« facile »*

In these examples, MWEs pose challenge to the translation system, which translates them word-by-word and does not consider them as a single token, which results in nonsense.

In this article, we first explain our method for MWEs extraction (automatically or semi automatically), then describe the specific corpora made for evaluating MT of MWEs, and finally propose a preliminary approach to handle MWEs in MT.

2 Building a Specific Corpus for Evaluating MT of MWEs

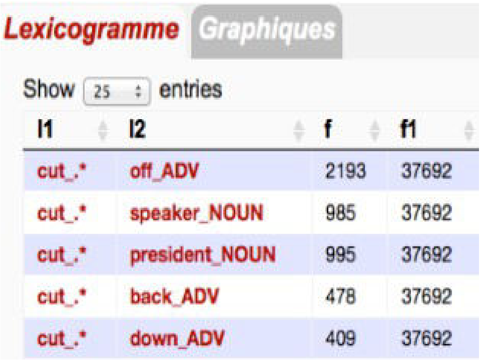
To assess and deal with these expressions in MT systems, it is necessary to have a system containing a lot of MWEs. Using the EmoConc[1] concordancer we can make complex searches in order to identify non-contiguous multiword expressions. Since the Emolex[2] corpus is a literary one, we have enriched it with the Europarl[3] (Koehn,

[1] http://emolex.u-grenoble3.fr/emoConc/index.php
[2] http://www.emolex.eu/
[3] http://www.statmt.org/europarl

2005),Ted[4] and News_commentary[5] corpora, in order to have a corpus with broader coverage.

2.1 Semi Automatic Tool for Extracting MWEs

We made a list of MWEs (Phrasal Verbs and idioms), based on electronic dictionaries available on the Net, Forums, etc. Moreover, we used the EmoConc *lexicogramme* tool (Kraif and Diwersy, 2012) to extract collocations, choosing a precise node-word and possibly a grammatical category as well as syntactic relations. For instance, to extract a PV based on the verb "*cut*" we retained only adverbs and prepositions. In Figure 1, for example, we can extract the following PVs: *cut off, cut back*, *cut down* and put them in our list.



l1	l2	f	f1
cut_.*	off_ADV	2193	37692
cut_.*	speaker_NOUN	985	37692
cut_.*	president_NOUN	995	37692
cut_.*	back_ADV	478	37692
cut_.*	down_ADV	409	37692

Figure1 : Extraction of *phrasal verbs* (using *Lexicogramme* tool*)*

2.2 MWEs Selection

Starting from the MWE list, we made queries on EmoConc database – looking for occurrences with distances (Verb-Prep) varying from 0 to 5, in the case of phrasal verbs. For each query, we first checked that the number of found occurrences was greater than 5. Then, we manually selected sentences containing valid MWEs. Afterwards, we moved on to automatic translation, using Google Translate. Whenever the system yielded a mistranslation of the MWEs, we kept the sentence, in order to have a challenging corpus. At the end of the selection process, we have retained between 5 to 10 sentences containing the same MWE.

Example of a query for the PV "*set up"* with a distance that varies between 0 and 5 follows:

<l=**set**,#2><>{0,5}<l=**up**,#1>::(.*,2,1)

2.3 En-Fr Corpus

After the process described above, the English-French test corpus contains 500 pairs of sentences extracted from different corpora (50,6% of the corpus from *Emolex*, 47% from *Europarl* and 2,4% from *News*), with 40 different PV (73.2% of the test corpus – named *pv366*) and 74 different idiomatic expressions (26.8% of the corpus – named *Idioms*).

3 Handling MWEs in MT

3.1 LIG (Moses) baseline Vs Google-TR

In order to determine the impact of our specific corpus (Tst-MWE) on MT performance, we made a *control* corpus of 500 random sentences, which has the same sampling (from the various sources) as the *challenging* test corpus.

System	Corpus of 500 sentences		PV in Tst-MWE	Idioms in Tst-MWE
	Control	Tst-MWE		
Moses Baseline	24.87%	20.83%	22.72%	15.21%
Google-TR	19,27%	18,97 %	18.67%	19.75%

Table 1:BLEU scores of our Moses Baseline vs Google-TR on the *challenging* (Tst-MWE) and *control* corpora

Table 1 reports BLEU scores of our Moses Baseline vs Google-TR on the *challenging* (Tst-MWE) and *control* corpora. These results show that our baseline Moses system described in (Besacier et al., 2012) is, all in all, more efficient on this corpus than Google-TR, especially in terms of PV translation.

Furthermore, the BLEU score (Papineni et al., 2002) of the control corpus is higher than that of the test corpus (20.83%). This confirms that the test corpus is more challenging.

However, Google-TR translates idioms better. Maybe, one explanation is that it may use specific dictionaries for idiomatic expressions.

3.2 Pre-processing of PVs for En-Fr SMT training

The main source of knowledge of the decoder is its phrase table. Indeed, the decoder consults this table in order to decide how to translate a source sentence into the target language. However, due to the automatic alignment errors of certain

[4] http://www.ted.com/

[5] http://www.statmt.org/wmt13/translation-task.html#download

words, extracted segments may not necessarily correspond to their source segments. Thus, to improve the alignment, we considered every PV as a single lexical unit, in order to force the segmentation during the alignment. For this purpose, the automatic identification of PV sequences is required. We conducted a parsing on the test corpus and the training corpus using *XIP* parser *(Aït-Mokhtar et al., 2002)* to get linguistic annotation for each form.

Then, using the parts of speech and some dependencies (as *NUCL_PARTICLE, MOD_POST*) provided by *XIP*, we adapted the output of the parser to get an *XML* version of the corpus compatible with the Moses toolbox, with an additional attribute "MWE = ***'verb-id, particle-id'***" for the PVs of our test corpus. Then we merged the verb with its particle (as a verb-particle compound), as in the following example:

keep your voice down, Hermione begged him.

keep-down your voice, Hermione begged him.

This approach has been applied to both test and training corpora (to train a new system called *Exp-MWE*).

To evaluate its usefulness, we calculated the BLEU scores for the outputs of the baseline system as well as *Exp-MWE*. In addition, in order to maximize the BLEU score, we used the MERT program (Minimum Error Rate Training) (Och, 2003), which allows us to adjust the weight of the different models involved in the translation process (such as language model and translation model).

System	**corpus of 500 sentences**		**PV in Tst-MWE**	**Idioms in Tst-MWE**
	Control	Tst-MWE		
Moses-Baseline	26.46%	23.14%	23.47%	22.03%
Exp-MWE	26.28%	23.68%	24.16%	21.83 %

Table 2: Baseline vs PV pre-processing (BLEU) – both systems optimized using MERT

The Exp-MWE system yields a slight improvement (+0.54% BLEU) in the whole corpus and (+0.69% BLEU) in the part corresponding to PV in Tst-MWE. This evaluation is somehow limited since BLEU is calculated on the full corpus while PV correspond to small events inside sentences. Thus an evaluation metric that could focus on MWEs only would be necessary to better evaluate the contribution of our approach.

The table also shows a BLEU decrease of 0.18% for the *control* corpus and of 0.30% for the *idioms* part of Tst-MWE.

Example of Baseline and Exp-MWE systems outputs hypothesis follows:

Source	surely they must ***call*** the operation ***off*** now ?
Reference	maintenant , ils doivent sûrement ***annuler*** l' opération .
Hyp (baseline)	ils doivent ***appeler*** l' opération maintenant ?
Hyp (+preproc)	ils doivent ***annuler*** le fonctionnement maintenant ?

3.3 Processing of idioms

We handled idioms using a method of constrained decoding available in the Moses decoder (Koehn, 2014). Using our idiom list, we developed a tool to identify idioms in our English source sentence and to put the correct translation of each expression between XML tags used by the decoder.

Example :
<idiom translation="***facile***"> ***piece of cake***
</idiom>

System	**Idioms in Tst-MWE**	**Overall corpus of 500 sent.**
Moses Baseline	15.21%	20.83%
Constrained decoding	30.71%	24.77%

Table 3: Baseline vs idiom pre-processing (BLEU)

The table 3 shows a significant increase in the BLEU score (15%) when evaluating on *Idioms* only. As far as the full corpus (500 sentences) is concerned, we achieve an overall BLEU improvement of 4 points.

4 Conclusion and Perspectives

In this paper, we described a specific bilingual corpus made of sentences that contain phrasal verbs or idioms. This corpus, resulting from a selection of sentences that were mistranslated by

Google-TR, appears to be challenging for our system as well, even if it outperforms Google-TR on this specific corpus. It confirms our hypothesis that the frequency of MWEs in a corpus may have an influence on translation quality. Our approach for pre-processing PV sequences in translation systems yielded only a slight improvement in performance. However, we proposed an efficient method to handle known idioms, which achieved a significant BLEU improvement on our specific test corpus. But this method has some limitations, especially if an expression has several meanings (ambiguous expression) or if it does not appear in the pre-existing dictionary of idioms (out-of-vocabulary expressions). Future works will try to address these limitations.

References

Papineni, K., Roukos, S., Ward, T., and Zhu W. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, p., July 2002, Philadelphia:311–318.

Och, F. 2003. Minimum Error Rate Training in Statistical Machine Translation, A*CL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics,* Vol 1, USA:160–167.

Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation, *Machine Translation***,** *MT Summit*, University of Edinburgh, Scotland.

Kraif, O. and Diwersy, S.. 2012. Le Lexicoscope: un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques, *Actes de la conférence TALN 2012*, Grenoble France: 399–406.

Besacier, L., Lecouteux, B., Azouzi, M., Quang, L. 2012. The LIG English to French Machine Translation System for IWSLT 2012**,** *Proceedings of the 9th International Workshop on Spoken Language Translation* (IWSLT).

Aït-mokhtar S., Chanod J. and C. Roux, 2002. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*,Vol-8, Issue 2-3, Cambridge University Press:121–144.

Koehn, P., 2014. *Statistical Machine Translation, System User Manual and Code Guide*, Université of Edinburgh Royaume-Uni.

Analysis of MultiWord Expression Translation Errors in Statistical Machine Translation

Natalia Klyueva
Charles University in Prague
Faculty of Mathematics and Physics
kljueva@ufal.mff.cuni.cz

Jeevanthi Liyanapathirana
Copenhagen Business School
Denmark
jl.ibc@cbs.dk

Abstract

In this paper, we analyse the usage of multiword expressions (MWE) in Statistical Machine Translation (SMT). We exploit the Moses SMT toolkit to train models for French-English and Czech-Russian language pairs. For each language pair, two models were built: a baseline model without additional MWE data and the model enhanced with information on MWE. For the French-English pair, we tried three methods of introducing the MWE data. For Czech-Russian pair, we used just one method – adding automatically extracted data as a parallel corpus.

1 Introduction

In this paper, we exploit a statistical machine translation (SMT) system, Moses, training it for two language pairs to explore how it cope with multiword expression translation in different languages. We will experiment with Czech-Russian, English-French language pairs to make sure that our conclusions are as language-independent as possible.

The problem of MWE in the area of SMT is a well-studied topic, next, we will name a few works that are most relevant to our work.

(Bouamor et al., 2012) described the way to extract an MWE bilingual lexicon lexicon from a parallel corpus and integrated this resource into an SMT system.

In the paper (Ghoneim and Diab, 2013) authors divided MWEs into several groups according to their parts of speech. They adopted the approaches to integrating MWE into SMT as described in (Carpuat and Diab, 2010): static (MWE on the source side are grouped with underscores) and dynamic (including MWE information straight into phrase tables) integration.

In our work, we will use three simpler methods of integrating MWE.

The paper is structured as follows. After the introduction, in Section 2, we present the notion and a basic classification of MWE. Next, we briefly describe the SMT system we are working with - Moses (Section 3). In Section 4 we present three methods to integrate MWE into SMT pipeline and test them for French-English language pair. In Section 5 we applied the most successful method from the previous experiment for Czech-Russian SMT, but MWE data we use here are sufficiently larger than for the previous experiment. Finally, we conclude in Section 6.

2 MultiWord Expressions

MWEs present a sequence of words with non-compositional meaning, they differ from language to language and are highly idiosyncratic. Even for the related languages we can not be sure if the structure of MWE is similar or not to say nothing about typologically different languages.

We can distinguish several types of the multiword expressions based on their part of speech and function in a sentence: noun multiword expressions, auxiliary multiword expressions, light verbs, idioms.

- Noun multiword expressions Multi-word expressions in our test sets are mainly named entities (NE) or belong to domain specific terminology (e.g. English-French : *military coup* – 'coup d'etat'. They generally contain a noun and some other part of speech. Those terms and NEs get translated properly if they were seen in the training data.

- Auxiliary multiword expressions present mainly multiword prepositions (e.g. English-French *with regard to* – 'en ce qui concerne' and SMT also does not have a problem to handle them properly because their co-occurrence in the data is quite frequent and parts of an expression are not separated by other words.

- Light verb constructions (LVC) are generally formed by a verb and a noun where a verb does not bare its initial meaning, so that the whole construction takes the semantics of the noun. Some multiword verbs have identical component words in the languages (Czech: *hrát úlohu*, Russian: *igrat' rol'* – 'to play role', and some not (Czech: *dát smysl* – 'give sense' vs. Russian: *imet' smysl* – 'have sense'. Generally, multiword expressions are translated properly within SMT when an LVC presents an n-gram, but when a verb is separated from a noun, this LVC is often mistranslated.

- Idioms are MWEs that can include words of any part of speech and they generally bear a meaning that has very little to do with any component of MWE. Idiomatic constructions often present a challenge to MT systems because they might be equal in the languages (contain the same words), but that is not always the case. For example, the English idiom : *kick the bucket* will be translated into French as *casser sa pipe* (which is the literal meaning) in systems like Google Translate, whereas the real meaning or translation should be "mourir" , which means "to die" in English .

Multiword Expressions have a better chance to be handled properly within SMT than within Rule-Based MT if no explicit modeling of MWE was integrated into systems. If some MWE is frequently used in the training data or it is lexically fixed, it is more likely to be translated correctly.

3 SMT Moses

In our experiments, we exploited the toolkit Moses (Koehn et al., 2007), an open-source implementation of a phrase-based statistical translation system. **The Moses toolkit**[1] relies on and also includes several components for data preprocessing and MT evaluation, like GIZA++[2] involved in finding word alignment, the SRI Language Modeling or SRILM Toolkit,[3] implementation of model optimization (Minimum Error Rate Training, MERT) on a given development set of sentences.

4 English-French SMT

This section will describe the experiments we conducted in translating multiword expressions from French to English. The subsections will explain the process of extracting multiword expressions, the word alignment procedure, and the integration of the extracted information for the statistical machine translation system.

4.1 Multiword Expression extraction

The first step of our experiments was to extract monolingual multiword expressions from a corpus. Choosing the proper multiword expressions was quite tricky, depending on the available resources.

We used a method of extracting multiword expressions using a linguistic rule based approach. We determined some of the most common types of linguistic rules which would effectively constitute in a multiword expression (e.g. Noun-Adj, Adj-Noun, Noun-Noun). Altogether, we defined 10 rules. Once the rules are determined, we use these linguistic rules to extract the potential multiword expression from the corpora.

Once the potential candidates for multiword expressions are extracted, the most frequent candidates in the extracted set were considered potential candidates to be used in training the machine translation system. In order to remove the irrelevant candidates in the process, we conducted a simple approach : if an MWE is included inside another, having the same frequency, we remove the one smaller in size. If not, we keep both.

We conducted this experiment to extract the multiword expression candidates in the French side of the corpus.

4.2 Word level alignment

Once the potential MWE s are extracted, the next step is to find the potential translations in English for them. For this purpose, we used the GIZA++ alignment toolkit. A parallel corpus (which included the MWEs we extracted) was trained, and

[1] http://www.statmt.org/moses/

[2] http://www.fjoch.com/GIZA++.html

[3] http://www.speech.sri.com/projects/srilm/

the alignments for the extracted MWEs were extracted out of the alignment output.

This way, the parallel MWE pairs were extracted out of the corpus. The next step was to incorporate that knowledge into a machine translation system.

4.3 Integrating information into Moses system

In order to integrate the above mentioned MWE pairs to the system, we conducted three different approaches.

4.3.1 Adding MWE pairs into training data

The first approach was based on simply including the extracted MWE pairs to the SMT system. This way, the extracted MWEs were considered as more training data.

4.3.2 Adding MWE pairs into the phrase table

In this approach, we made use of the phrase table which is created in the Moses SMT system. We inserted the extracted MWE pairs as phrase pairs in the lexical table which is generated while training the MT system. The probability for the lexical phrase pair (which is, here, a MWE pair) is set to 1.

4.3.3 Integrating features into Moses decoder

In the third approach, we inserted a simple feature to the Moses feature file, and used it for the MERT training. This feature simply mentions whether the phrase pair in concern is a multiword expression or not.

4.4 Experiments

As mentioned earlier, we use Moses as our statistical machine translation system. In order to extract the linguistics features, we used Stanford parser, and the TreeTagger toolkit. Plus, to generate the alignment model (to extract the MWE pairs), we used GIZA++ toolkit.

To conduct this experiment, we extracted 50 potential MWE candidates. Then, we conduct the above mentioned approaches for English to French data sets. We consider the Europarl parallel corpus for French to English for this purpose.

Table 1 shows the dataset we used for training the statistical machine translation system.

Table 2 below shows the BLEU scores we got for a test set of 10000 sentences , which include the MWEs we extracted. The baseline approach depicts the normal BLEU score we get for the parallel corpus, and the next three lines demonstrate the BLEU score we obtained using each of the approaches mentioned in section 4.3.

	French	English
Sentences	32000	33000
Words	120000	150000

Table 1: Europarl corpus : French to English . The statistics show the number of words and sentences in the corpus in each side

Method	BLEU
Baseline System	21.67
Adding MWE pairs into training data	21.88
Adding MWE pairs into the phrase table	21.68
Integrating features into Moses decoder	19.2

Table 2: BLEU Scores for each approach

Table 2 shows that two approaches we conducted slightly increase the BLEU score. However, the approach of integrating features into Moses decoder degrades the performance. This gives a positive potential to the fact that incorporating MWE s to the SMT system in different manners can effectively increase the BLEU score.

It should be also mentioned that it is quite difficult to evaluate the efficiency of our proposed approaches by incorporating a significantly small number of MWEs, e.g. 50. Also, the alignment models can also give some amount of noise in their alignments, so the extracted MWE pairs are not 100% accurate. These reasons might have contributed to the fact of having a relatively low increase in BLEU score.

5 Czech-Russian SMT

Our second experiment with Czech-Russian language pair includes only one method of introducing MWE. We will exploit the simplest method described in Section 4.3.1 - adding MWE lexicon as a parallel corpus and retraining the system on the enhanced data.

5.1 Baseline SMT

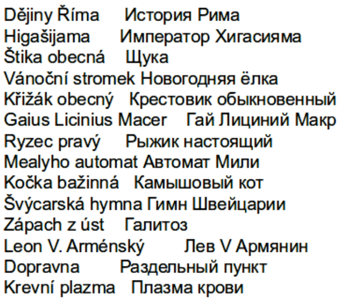
We trained a baseline system on data coming from news domain[4] and from the domain of fiction[5]. Europarl corpus does not include version in Russian, so we can not add parallel data from this resource. The data for training a language model for the target language - Russian - were compiled from various online resources, see (Bílek, 2014) for details. Table 3 presents the statistics of the training data.

corpus	sentences
news	93432
fiction	148810
total	242242

Table 3: Size of training data

5.2 MWE from wikipedia headlines

We used a list of names and phrases from Wikipedia headlines for the pair Czech-Russian. The headlines were automatically extracted from the wikipedia dumps in XML (`https://dumps.wikimedia.org/`). The headlines were not necessarily multiword expressions, but for the sake of our experiment, we extracted only MWEs. Following is the example of several entities from the list:



Dějiny Říma	История Рима
Higašijama	Император Хигасияма
Štika obecná	Щука
Vánoční stromek	Новогодняя ёлка
Křižák obecný	Крестовик обыкновенный
Gaius Licinius Macer	Гай Лициний Макр
Ryzec pravý	Рыжик настоящий
Mealyho automat	Автомат Мили
Kočka bažinná	Камышовый кот
Švýcarská hymna	Гимн Швейцарии
Zápach z úst	Галитоз
Leon V. Arménský	Лев V Армянин
Dopravna	Раздельный пункт
Krevní plazma	Плазма крови

Figure 1: Czech-Russian MWEs from Wikipedia headlines

The automatically extracted data are not very clean; there are no light verb constructions and hardly any idioms, mostly they are Named Entities. Total number of MWE pairs extracted from the Wikipedia is 87354.

5.3 Results of the experiment

Using the factored configuration of Moses, we ran two experiments:

- the baseline with models trained on data without the Wikipedia headlines
- model trained on data including the headlines

Table 4 demonstrates the difference in performance between the baseline system and the system trained on data with additional MWE resource. In addition to BLEU, we calculated the number of out-of-vocabulary (OOV) words - searching for Latin characters in the translation output (Czech words left untranslated by Moses).

	BLEU	OOV
Baseline system	17,23%	1216
With MWE	17,90%	1011

Table 4: BLEU score and OOV rate for SMT trained on data with and without MWE resource

The BLEU score in the second experiment was slightly better than in the baseline, but, evidently, this improvement is insignificant. The number of out-of-vocabulary words decreased by 205 individual tokens. This may be attributed to the positive effect of adding new data.

5.4 Examples of improved MWE

We examined the list of OOV words in the output from the two experiments. Among those 205 words that were recognized and translated in the second experiment, there were MWEs from the added resource, such as *Carlo Ancelotti*, *Amschel Rothschild*, *alt soprán* etc. The following MWEs were not translated or mistranslated in baseline, but were translated correctly according to the added data in the improved setup: *Higgsův boson* – 'Bozon Higgsa' (Higgs boson), *Velký hadronový urychlovač* – 'Bol'shoy adronniy collajder' (LHC), *Praní špinavých peněz* – 'Otmivanie deneg' (money laundering) etc.

6 Conclusion

In this work, we presented experiments with integrating MWE into SMT for the two language pairs - French-English and Czech-Russian. We tested three methods of including MWE information into SMT. It turned out that for the concrete language pair (French-English) and the concrete MWE list

[4] `http://ufal.mff.cuni.cz/umc/cer/`

[5] Czech-Russian side of Intercorp, `https://ucnk.ff.cuni.cz/intercorp/`, not an open-source

the method of introducing MWE as additional parallel data scored better than other methods. We adopted this method for the pair Czech-Russian and added an automatically extracted resource. In both cases, the increase in BLEU score was very little, but this often happens when improving concerns one concrete linguistic issue.

Acknowledgments

This work has been using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

References

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In *LREC*, pages 674–679.

Karel Bílek. 2014. A Comparison of Methods of Czech-to-Russian Machine Translation. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Praha, Czechia.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245. Association for Computational Linguistics.

Mahmoud Ghoneim and Mona T. Diab. 2013. Multiword Expressions in the Context of Statistical Machine Translation. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 1181–1187.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.