Computational and Corpus-based Phraseology
Recent advances and interdisciplinary approaches

EUROPHRAS 2017
London, 13-14 November 2017

Proceedings of The 3rd Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2017)

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

14 November 2017

ISBN 978-2-9701095-4-9

9 782970 109549

2018. Editions Tradulex, Geneva

ISBN: 978-2-9701095-4-9

This document is downloadable from www.tradulex.com

Workshop Chairs

Gloria Corpas Pastor (Universidad de Málaga, Spain)
Ruslan Mitkov (University of Wolverhampton, United Kingdom)
Johanna Monti (Università degli Studi di Napoli "L'Orientale", Italy)
Violeta Seretan (Université de Genève, Switzerland)

Programme Committee

Iñaki Alegria (University of the Basque Country)
Giuseppe Attardi (University of Pisa)
Philippe Blache (Aix-Marseille University)
Fabienne Cap (Uppsala University)
Matthieu Constant (Université de Lorraine)
Antoine Doucet (University of La Rochelle)
Thomas François (Université catholique de Louvain)
Philipp Koehn (Johns Hopkins University)
Valia Kordoni (Humboldt-Universität zu Berlin)
Michael Oakes (University of Wolverhampton)
Carla Parra Escartín (ADAPT Centre, Dublin City University)
Pavel Pecina (Charles University)
Carlos Ramisch (Aix Marseille University)
Agata Savary (Université François Rabelais Tours)
Gerold Schneider (University of Zurich)
Max Silberztein (University of Franche-Comté, Besançon)
Kathrin Steyer (Institut für Deutsche Sprache, Mannheim)
Amalia Todirascu (University of Strasbourg)
Beata Trawinski (Institut für Deutsche Sprache, Mannheim)
Agnès Tutin (Université Grenoble Alpes)
Aline Villavicencio (Federal University of Rio Grande do Sul)
Veronika Vincze (Hungarian Academy of Sciences)
Martin Volk (University of Zurich)
Andy Way (ADAPT Centre, Dublin City University)
Mike Rosner (University of Malta)

Invited speaker

Carlos Ramisch

Carlos Ramisch is an assistant professor in computer science at Aix Marseille University and a researcher in computational linguistics at Laboratoire d'Informatique Fondamentale de Marseille (France). He has a double PhD in Computer Science from the University of Grenoble (France) and from the Federal University of Rio Grande do Sul (Brazil). He is passionate about languages, and in particular about multiword expressions. His long-term research goal is integrating multiword expressions processing into NLP applications. He is interested in MWE discovery, identification, representation and translation, lexical resources, machine translation, computational semantics and machine learning. He was co-chair of the 2010, 2011, 2013 and 2017 editions of the MWE workshop, area chair for MWEs in *SEM 2012, guest editor of the 2013 special issue on MWEs of the ACM TSLP journal, active member of the PARSEME community, one of the organizers of the PARSEME shared tasks, member of the standing committee of the SIGLEX MWE-Section, author of a book on MWE processing, local coordinator of the ANR PARSEME-FR project, and developer of the mwetoolkit, a free tool for automatic MWE processing.

Abstract

Putting the Horses before the Cart: Identifying Multiword Expressions before Translation

Translating multiword expressions (MWEs) is notoriously difficult. Part of the challenge stems from the analysis of non-compositional expressions in source texts, preventing literal translation. Therefore, before translating them, it is crucial to locate MWEs in the source text. We would be putting the cart before the horses if we tried to translate MWEs before ensuring that they are correctly identified in the source text. This paper discusses the current state of affairs in automatic MWE identification, covering rule-based methods and sequence taggers. While MWE identification is not a solved problem, significant advances have been made in the recent years. Hence, we can hope that MWE identification can be integrated into MT in the near future, thus avoiding clumsy translations that have often been mocked and used to motivate the urgent need for better MWE processing.

MUMTTT 2017 Academic Programme

Tuesday, 14th November 2017

Session 1: ***Multi-words in Machine Translation***

Multi-word Adverbs - How well are they handled in Parsing and Machine Translation?
Martin Volk and Johannes Graën

Chinese Separable Words in SMT
Gongbo Tang and Fabienne Cap

Out of the tailor or still in the woods? An empirical study of MWEs in MT
Fabienne Cap

Tuesday, 14th November 2017

Invited Talk

Putting the Horses before the Cart: Identifying Multiword Expressions before Translation
Carlos Ramisch

Tuesday, 14th November 2017: 17:30 - 18:30

Session 2: ***Lexical and morphological aspects in MWU representation***

The Corpus Analysis of Path Elements of the Verb otići/oditi ‘leave’ in Croatian and Slovene
Goranka Blagus Bartolec and Ivana Matas Ivanković

Morphology of MWU in Quechua
Maximiliano Duran

Table of Contents

Preface by the workshop Chairs

This volume documents the proceedings of the 3rd Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2017), held on 4 November 2017 as part of the EUROPHRAS 2017 conference: "Computational and Corpus-based Approaches to Phraseology: Recent advances and interdisciplinary approaches" (London, 13-14 November 2015), jointly organised by the European Association for Phraseology (EUROPHRAS), the University of Wolverhampton (Research Institute of Information and Language Processing) and the Association for Computational Linguistics – Bulgaria. The workshop was held 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 Ruslan Mitkov (University of Wolverhampton), Johanna Monti (Università degli Studi di Sassari), Gloria Corpas Pastor (Universidad de Málaga) and Violeta Seretan (Université de Genève).

The topic of the workshop was the integration of multi-word units in machine translation and translation technology tools. In spite of the relative progress achieved for particular types of units such as verb-particle constructions, the identification, interpretation and translation of multi-word units in general still represent open challenges, both from a theoretical and a practical point of view. The idiosyncratic morpho-syntactic, semantic and translational properties of multi-word units pose many obstacles even to human translators, mainly because of intrinsic ambiguities, structural and lexical asymmetries between languages, and, finally, cultural differences. The aim of the workshop was to bring together researchers and practitioners working on MWU processing from various perspectives, in order to enable cross fertilisation and foster the creation of innovative solutions that can only arise from interdisciplinary collaborations. The present edition of the workshop provided a forum for researchers and practitioners in the fields of (Computational) Linguistics, (Computational) Phraseology, Translation Studies and Translation Technology to discuss recent advances in the area of multi-word unit processing and to coordinate research efforts across disciplines in order to improve the integration of multi-word units in machine translation and translation technology tools. The programme included 5 oral presentations, and featured an invited talk by Carlos Ramisch, Aix-Marseille University, France. 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. Finally, we would like to thank the local organisers for all their work and their effort in the organisation of the workshop.

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

Multi-word Adverbs – How well are they handled in Parsing and Machine Translation?

Martin Volk, Johannes Graën
University of Zurich
Institute of Computational Linguistics
volk|graen@cl.uzh.ch

Abstract

Multi-word expressions are often considered problematic for parsing or other tasks in natural language processing. In this paper we investigate a specific type of multi-word expressions: binomial adverbs. These adverbs follow the pattern *adverb + conjunction + adverb*. We identify and evaluate binomial adverbs in English, German and Swedish. We compute their degree of idiomaticity with an ordering test and with a mutual information score. We show that these idiomaticity measures point us to a number of fixed multi-word expressions which are often mis-tagged and mis-parsed. Interestingly, a second evaluation shows that state-of-the-art machine translation handles them well – with some exceptions.

1 Introduction

We work on the annotation of large corpora for linguistic research and information extraction. We noticed that multi-word adverbs often cause confusion to the PoS tagger and subsequently to the parser and thus require special treatment. We investigate a specific type of multi-word expressions: binomial adverbs. These adverbs follow the pattern *adverb + conjunction + adverb*. English examples are *by and large, first and foremost, over and over*. In German we find *ab und zu, ganz und gar, nach wie vor* (EN: occasionally, completely, still). The most prominent example in Swedish is *till och med*, but there are many others like *blott och bart, helt och hållet, om och om (igen)* (EN: purely and simply, completely, again and again). We searched manually annotated corpora for English, German and Swedish for occurrences of such binomial adverbs. We also examined an automatically annotated version of Europarl for these three languages. We found that German and Swedish have more occurrences of binomial adverbs in both the manually annotated corpora and the automatically tagged and parsed Europarl. We will present the comparison across the three languages in sections 2.2 and 2.3.

We will start with a definition of binomial adverbs as a subclass of multi-word adverbs. Section 3 has our results of the parsing evaluation and section 4 describes our machine translation evaluation.

This paper is thus an evaluation paper and not a methodology paper. We identify a subclass of multi-word expressions that need special treatment in order to improve parsing results and machine translation (MT).

2 Multi-word Adverbs

In this paper we focus on a subclass of multi-word adverbs which we call **binomial adverbs** since they intersect with binomials as described by Gereon Müller (1997) and Sandra Mollin (2014). This class of adverbs is interesting since the “adverbs” that make up the construction are often taken from other parts-of-speech. For example, the German binomial adverb *nach und nach* (EN: gradually) is constructed out of conjoined prepositions. The frequent Swedish binomial adverb *till och med* (EN: even, through) is annotated in the Stockholm-Umeå Corpus (SUC) as conjoined adverbs 120 times and as conjoined prepositions 16 times. Certainly the “adverbs” *till* and *med* in this binomial construction are much more frequently used as prepositions and verb particles than as adverbs (see table 2). Sometimes other parts-of-speech are used as e.g. the adjective in English *by and large* or the Swedish noun *hållet* (EN: distance, direction) in *helt och hållet*. These PoS am-

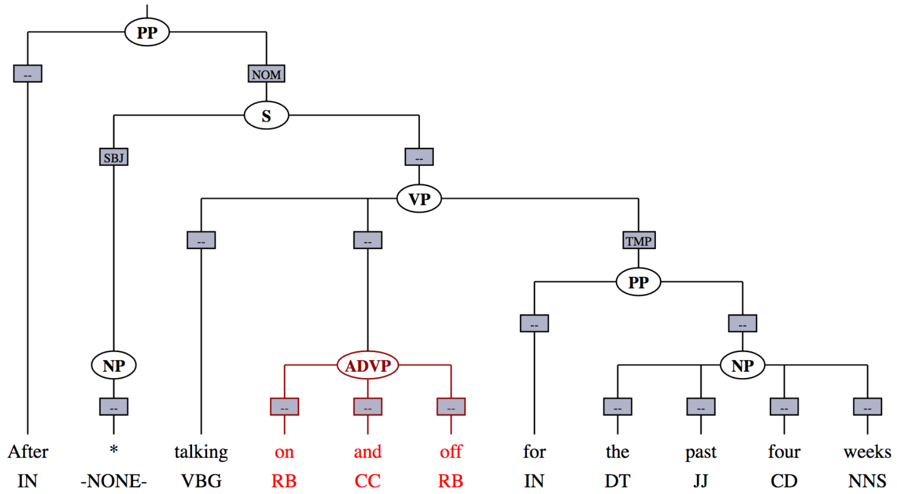


Figure 1: Syntax tree with multi-word adverb (*on and off*) from the Penn treebank. The multi-word adverb is annotated as adverbial phrase (ADVP).

biguities may lead to processing errors in parsing or translation.

2.1 Related work

Binomial constructions have been studied in linguistics for many years. (Bendz, 1965) is an early monograph on “word pairs” in Swedish with comparisons to Danish, English and German. Bendz deals not only with adverbs but all kinds of coordinated word pairs. He presents a semantic classification to distinguish the binomials into

1. opposition pairs (e.g. EN: *sooner or later, to and fro*, DE: *dick und dünn, weit und breit*, SV: *tjockt och tunt, vitt och brett*)
2. enumeration pairs (e.g. SV: *män, kvinnor och barn; tid och rum*), but no examples for adverbs
3. synonym pairs (e.g. EN: *first and foremost, simply and solely*, DE: *frank und frei, ganz und gar*, SV: *blott och bart, helt och hållet*)

Bendz discusses many aspects of word pairs such as inheritance of the constructions from Latin, their prominence in the literature, but also formal properties such as alliteration, assonance and rhyme. He mentions word order constraints and formulates the hypothesis that “the more frequent a word pair becomes, the fixer is word order” (p. 21, translation by the authors). He links the ordering tendencies of unequal words to Behaghel’s law of increasing terms which predicts that shorter words come before longer words. Some examples from English might illustrate this law: *first and foremost, *foremost and first, far and away, *away and far, now and again, *again and now*.

Müller (1997) presents a detailed study of binomial constructions in German (e.g. *Fug und Recht, samt und sonders*) which includes binomial adverbs. He is particularly interested in order constraints which he regards as a defining feature of binomial constructions.

Müller elaborates that end rhyme, alliteration (*ganz und gar*) and assonances (the repetition of vowel sounds to create internal rhyming; e.g. *dann und wann*) are typical properties of binomial constructions.

In a recent book Mollin (2014) discusses ordering constraints of English binomials in great detail. Mollin performs a study on the British National Corpus (BNC), a 100 million word corpus collected in the 1990s. She relies on automatic PoS tagging and searches for pairs of the same parts-of-speech like ‘noun *and* noun’, ‘verb *and* verb’, ‘adverb *and* adverb’ etc. She investigated all candidates where the more frequent sequence occurs 50 times or more. This resulted in 544 types. Among these, there are 20 adverbs, with *up and down, here and there, now and then* being the most frequent ones. Mollin judged the candidates by a so called irreversibility score which is the ratio of the more frequent order against the frequency of both orders. The highest irreversibility scores of 100% go to *back and forth, out and about, today and tomorrow* all of which were only found in the BNC in the given sequence.

There have been a lot of studies on the automatic parsing and translation of multi-word expressions. But surprisingly few of them deal with multi-word adverbs or binomials. Widdows and Dorow (2005) find idiomatic expressions of the form 'noun *and/or* noun' in the BNC by exploiting ordering constraints. They elaborate on the difference between symmetric and asymmetric relationships between nouns, where asymmetric relations (one order being clearly more frequent than the other) may indicate idiomaticity or a number of other constraints such as hierarchies, gender asymmetries or temporal order. Michelbacher et al. (2007) explore the properties of asymmetric association measures which pay tribute to the, oftentimes, asymmetric nature of collocations.

Volk et al. (2016) have investigated multi-word adverbs for German and their impact on PoS tagging accuracy and the re-combination of separated verb prefixes to their respective verbs. Since some separated verb prefixes are homographs with prepositions (and obviously derived from prepositions) and also used in binomial adverbs (as e.g. *ab und zu, nach und nach*), it is important to identify the binomial adverbs in order to avoid confusion with separated verb prefixes and to prevent subsequent erroneous verb lemmas and syntax structures.

Out of the large pool of NLP studies on multi-word expressions in general let us mention Ramisch (2015) who introduces methods for the discovery of multi-word expressions, among others the computation of collocation scores. Constant and Nivre (2016) show how to integrate MWEs into dependency parsing.

Nasr et al. (2015) propose a method for jointly tokenizing and parsing adverb-conjunction combinations in French (*ainsi que, bien que*). Their problem is similar to ours in that the combination shows no internal variability which makes it easy to spot but is ambiguous because of cases with literal usage (e.g. *ainsi que* is reported to be a multiword unit in only 76.6% of the cases). Their method relies on subcategorization information for the verbs with respect to the verbs' tendency to take subordinate *que* clauses. They report on clear improvements in parsing these cases.

From the long history of MWEs in machine translation we refer to two recent works. Bouamor et al. (2012) show how to integrate MWEs into statistical machine translation for French-English. Tan and Pal (2014) extract MWEs for Hindi and English, integrate them into MT and report on an improvement in MT quality.

2.2 Binomial Adverbs in Manually Curated Corpora

In order to get an overview of the frequency of binomial adverbs we evaluated the Penn Treebank for English, the TIGER treebank for German, and SUC for Swedish (all of which have roughly 1 million tokens).

In the German TIGER treebank we find 211 syntactic constituents labeled as coordinated adverb phrase (CAVP), corresponding to 110 types. The top frequent ones are *nach wie vor* (66 occurrences), *mehr oder weniger* (10), and *nach und nach* (7). Only 26 types occur more than once.

For Swedish we used the Stockholm-Umeå Corpus (SUC) which is a manually checked corpus with lemmas and parts-of-speech. SUC is part of the Swedish treebank. In SUC we find 985 sequences *adverb + conjunction + adverb* which are potential candidates for binomial adverbs. *till och med* is the most frequent sequence with 120 occurrences, followed by *mer eller mindre* (61) and *då och då* (45 occurrences). The prominence of *till och med* is underlined by the fact that it, additionally, occurs in SUC as an acronym with 52 occurrences (spelled as *t o m* or *t.o.m.*). Furthermore *till och med* is in SUC 16 times as a conjunction of prepositions. So, it is truly ambiguous with a clear frequency bias towards being a binomial adverb.

If we broaden the search for "adverb or particle or preposition" in the conjunction pattern, then we get 1510 hits in SUC. An analogous query in the Penn Treebank results in only 238 hits, which gives a first indication that binomial adverbs are more frequent in Swedish than in English. The most frequent hits in the Penn treebank are *up and down* (13), *in and out* (8), and *sooner or later, back and forth* (7 each). Given the syntactic annotation we can constrain our search to cases where the sequence has an adverb phrase (ADVP) as mother node which reduces our hits to 115. Figure 1 shows an example tree from the Penn Treebank with the binomial adverb *on and off.*

2.3 Binomial Adverbs in Europarl

We annotated Europarl in order to extract all candidates for binomial adverbs. We took a version of Europarl that has 43.1 million tokens for English, 41.1 million tokens for German and 36.1 million

	EN glosses	EN translation	SUC freq	SUC type
först och främst	first and mainly	first and foremost	12	coord adverb
helt och hållet	whole and distance	completely	12	coord adverb
i och för (sig)	in and for	in itself / actually	36	coord prep
i och med	in and with	since / as a result of	61	coord prep
kors och tvärs	cross and across	criss-cross	6	coord adverb
om och om (igen)	again and again	again and again	9	coord adverb
till och med	to and with	even / until	136	coord adv / prep

Table 1: Idiomatic Swedish binomial adverbs with words that have multiple PoS tags (particles, prepositions, nouns). Frequencies (ignoring case) are from the Stockholm Umeå Corpus (1.16 million tokens, mixed texts).

	preposition PP	adverb AB	particle PL	miscellaneous
för	11,035	401	63	101 KN, 28 SN, 44 VB
i	25,522	-	123	4 misc
med	11,063	166	544	2 misc
om	5011	143	400	1 KN, 2395 SN, 2 misc
till	9500	240	674	

Table 2: Part of Speech tag frequencies in SUC for particles that occur in multi-word adverbs (lower case usage only). Miscellaneous PoS tags include conjunction (KN), subjunction (SN), verb (VB).

tokens for Swedish. We tagged the English and German parts with TreeTagger and parsed with MaltParser. We annotated the Swedish part with a combination of Stagger und Maltparser with the standard model for Swedish[1]. All language-specific PoS tags were also mapped to universal PoS tags.

This allows us to extract binomial adverbs for the three languages with the same query. Since we know that adverbs in these constructions are sometimes confused with particles (PRT) and prepositions (ADP), we search for the pattern (ADV or ADP or PRT) followed by a conjunction followed by one of these three PoS tags again (ADV or ADP or PRT). This leads to the following results. For Swedish we find 37,973 occurrences (with 6983 binomial ADV types), while for English we have 23,509 occurrences (with 8034 types). So, given that the Swedish part of our Europarl corpus has 10% less tokens overall than English, this is a clear indication that binomial adverbs are more common in Swedish than in English.

The frequencies for German are not directly comparable with English and Swedish because of differences in the annotation of adjectives. German adjectives that function as adverbs or predicates are tagged as ADJD (in contrast with attributive adjectives which get the tag ADJA). If we were to map all ADJDs to ADVs in the universal PoS tag set, we would get way more adverbs than in English or Swedish. We therefore decided to skip the ADJD cases and use the same query for German as for the other two languages. This results in 19,427 occurrences (with 3830 binomial ADV types).

A closer look at Swedish binomial adverbs reveals that some of them are tagged as ADJ as well. For example, *helt och hållet* is tagged as "ADV CONJ ADV" 2448 times, but also 98 times erroneously as "ADJ CONJ ADJ", and 46 times as "ADJ CONJ ADV" (plus 18 times with other miscellaneous tag combinations).

The most frequent ones for Swedish are *till och med, först och främst, helt och hållet*[2], while for English they are *whether or not, once and for (all), more and more*. The top frequent German candidates are *nach wie vor, voll und ganz, (so) schnell wie möglich*. The examples show that sometimes

[1] http://maltparser.org/mco/swedish_parser/swemalt.html

[2] The Swedish Europarl has also 1096 occurrences of the acronym *t.o.m.* which stands for *till och med*.

we catch candidates that are part of larger idiomatic expressions (as in the case of *once and for (all); (so) schnell wie möglich* (EN: as fast as possible)).

2.4 Binomial Adverbs and Idiomaticity

The above sections exemplify that many binomial adverbs are true multi-word expressions (with non-compositional semantics) that need special treatment in natural language processing. In order to zoom in on idiomatic binomial adverbs, we used two methods. First we checked for order restrictions. If a candidate "X conjunction Y" has a certain corpus frequency but the opposite order "Y conjunction X" does not occur (or occurs with a much smaller frequency), then this increases the likelihood that the candidate is an idiomatic expression.

This check excludes English candidates like *clearly and fully* (3 times in either order), German candidates like *heute und jetzt* (6 times in either order), and Swedish candidates like *där och när* (4 times in either order), and *alltid och överallt* which occurs 13 times in this order and 7 times in the reverse order in Europarl. In this way we exclude several hundred candidates in each language.

Obviously this method does not work for candidates with word repetitions. Our Europarl search resulted in 47 such reduplication candidates in Swedish[3] with *mer och mer, då och då, om och om (igen), längre och längre, så och så* being the top frequent ones. It is striking that reduplications are often used with comparative forms as e.g. *bättre och bättre, snabbare och snabbare, mindre och mindre* and with words that already stand for repetitions *åter och åter, igen och igen, till och till* which are intensified in this way. The observations for English and German with respect to reduplications are very similar.

Second, we computed collocation scores (mutual information scores, MI) for all candidates "X conjunction Y". For this we used the pair frequencies of "X conjunction" and "conjunction Y" in comparison with the frequency of the triple "X conjunction Y". Our formula is

$$MI(X, C, Y) = \log_2 \frac{N \cdot f(X, C, Y)}{f(X, C) \cdot f(C, Y)}$$

with C being the conjunction between the particles X and Y, and N being the number of tokens in the corpus. In this way the MI score predicts the probability of "X conjunction" being followed by Y, and the likelihood of "conjunction Y" being preceded by X.

We set a minimum threshold of 6 occurrences for the triple. This cutoff results in 437 English candidate triples. High frequency candidates like *once and for, over and above, again and again* get high MI scores of 12 to 14. At the lower end we find *here and on, up and in* that are certainly not idiomatic multi-word units.

For German high MI examples are *hin oder her* (MI: 21) and *nie und nimmer* (MI: 20.8). Frequent candidates *ganz und gar, mehr und mehr, nach und nach* receive prominent MI scores between 14 and 16 which increases their likelihood of being multiword units.

For Swedish the MI scores leave 426 candidate triples in the game. The top frequency candidates *till och med, först och främst, helt och hållet* get high MI scores of 10 and above, with *kors och tvärs, blott och bart, sönder och samman* receiving scores of 20+. At the other end of the scale we are able to rule out candidates with scores below 8 such as *här och i, nu och då, snabbt och på.*

In conclusion, binomial adverbs cover the whole spectrum of idiomaticity and can only be interpreted correctly when their ordering constraints and their collocation strengths are appropriately considered. Following the above considerations we picked 7 candidates per language for evaluation in parsing and machine translation. Criteria for the selection were idiomaticity, frequency and PoS ambiguities. Tables 3, 4 and 5 show the selected candidates.

3 Evaluation of Binomial Adverbs in Parsing

For the evaluation of binomial adverbs we ran MaltParser over the English, German and Swedish parts of the Europarl corpus. We then profiled the parsing results of our selected binomial adverbs. We counted the PoS patterns assigned to these adverbs, their dependency patterns (ignoring the labels) and their dependency label patterns. For example, in figure 2 we find the PoS pattern "ADV - CONJ - ADV" for *on and on* which is the desired tag sequence. Both the conjunction and the second adverb are marked as being dependent on the first adverb. These are the desired dependencies in English and Swedish. The German dependency

[3] Such repetition candidates are called echoics in (Mollin, 2014).

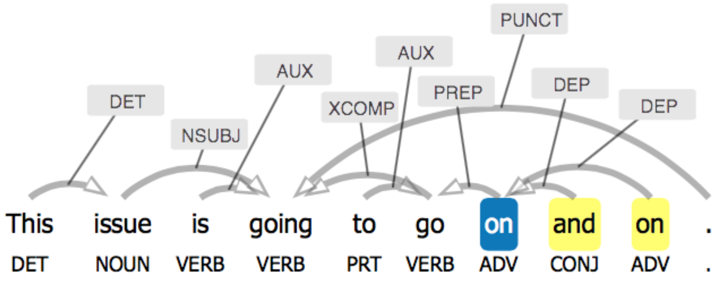

PUNCT
AUX
AUX
DET
NSUBJ
XCOMP
PREP
DEP
DEP
This issue is going to go on and on .
DET NOUN VERB VERB PRT VERB ADV CONJ ADV .

Figure 2: English sentence from Europarl with multi-word adverb (*on and on*) correctly tagged and parsed (disregarding the dependency labels).

	Europarl	PoS acc.	Parsing acc.	MT acc. EN → DE
by and large	161	0.6%	72.1%	100%
first and foremost	1496	17.2%	80.4%	100%
now and again	46	100%	100%	90.0%
on and off	20	55.0%	70.0%	32.0%
on and on	43	46.5%	81.4%	71.4%
out and about	7	85.7%	42.9%	66.7%
over and over	150	52.0%	50.0%	91.7%

Table 3: English binomial adverbs selected for evaluation.

parser marks the conjunction as dependent on the first adverb and the second adverb as dependent on the conjunction. When MaltParser assigns these dependencies, we count the parse of the binomial adverb as correct.

Table 3 shows the results for the English adverbs in our evaluation. For example, *by and large* occurs 161 times in Europarl (counting both upper and lower case occurrences). But only a single occurrence (0.6%) gets the PoS tags “ADV - CONJ - ADV”. Instead 145 occurrences are tagged as “ADP - CONJ - ADJ” (adposition - conjunction - adjective), and 15 occurrences are tagged as “ADP - CONJ - ADV”. Despite the high error rate in PoS tagging, the dependency arcs are correct in 72.1% (disregarding the dependency labels), indicating that this is a multi-word unit. This might be influenced by the fact that we parsed Europarl based on language-specific PoS tags and only later converted them into universal PoS tags.

Tagging the particles as adpositions also accounts for the 53.5% PoS errors with *on and on*. Still we observe a high accuracy of unlabeled dependencies with 81.4%. Figure 2 shows this adverb correctly parsed, while figure 3 has an incorrectly parsed example.

One should note that our evaluation may fall victim to cases of literal, i.e. non-idiomatic, usage. For example our extracted list includes the sentence ... *which is being worked **on and on** which we may make progress* .... But such occurrences of literal usage are rare.

Tables 4 and 5 have the corresponding results for the selected German and Swedish adverbs. We cannot discuss all observations here, but let us focus on the most prominent Swedish binomial adverb.

For *till och med* we expect to get a dependency profile where both *och* and *med* have a “head” dependency to *till*. And *till* then has a dependency as contrastive adverbial (CA), attitude adverbial (MA) or other adverbial (AA) or (seldom) as time adverbial (TA) to the appropriate word in the sentence. The good news is that *till och med* has the desired dependencies in 95.5% of the cases in Europarl.

It is also positive that *till* has plausible dependency labels in about 70% of the cases. However, this leaves 30% of the cases with dubious external dependency labels. For example, there are 4 occurrences of *till och med* with *och* being the root (!) of the sentence. All 4 occurrences are immediately preceded by the conjunction “eller”. There are 162 occurrences where *till* is the root of the sentence which is also unlikely. There are over one thousand occurrences with *till* labeled as time

	EN translation	Europarl	PoS acc.	Parsing acc.	MT acc. DE → EN
ab und an	occasionally	19	0%	100%	80.0%
ab und zu	occasionally	65	0%	98.5%	83.9%
eh und je	ever	24	0%	100%	90.0%
hin und wieder	sometimes	84	98.90%	85.7%	67.7%
kreuz und quer	criss-cross	20	0%	100%	55.0%
nach und nach	gradually	373	85.0%	98.7%	87.1%
nach wie vor	still	4723	99.9%	71.3%	96.8%

Table 4: German binomial adverbs selected for evaluation.

	EN translation	Europarl	PoS acc.	Parsing acc.	MT acc. SV → EN
först och främst	first and foremost	3988	99.4%	92.0%	95.8%
helt och hållet	completely	2610	93.8%	40.0%	100%
i och för (sig)	in itself / actually	249	52.2%	51.4%	48.0%
i och med	since / as / with	3035	67.9%	67.6%	81.8%
kors och tvärs	criss-cross	21	100%	14.3%	33.3%
om och om	again and again	250	100%	54.0%	100%
till och med	even / until	6802	99.9%	95.5%	94.3%

Table 5: Swedish binomial adverbs and conjunction (*i och med*) selected for evaluation.

adverbial which are mostly wrong. So, this calls for a special treatment of the binomial adverbs either prior to parsing or during parsing.

4 Evaluation of Binomial Adverbs in Machine Translation

We also evaluated how well a state-of-the-art MT system, Google Translate, handles the selected binomial adverbs. For each adverb we extracted sentences from various corpora. We did not use Europarl sentences for this evaluation since chances are high that this corpus is part of the training data for Google Translate. We sorted the extracted sentences by length (number of tokens) and dropped the short ones (less than 10 tokens) because they may not provide enough context for MT, and we dropped the long ones (more than 50 tokens) because they may confuse the MT system and because they make manual evaluation more time-consuming. From the remaining sentences we selected 30 per adverb and fed them to Google Translate. We translated EN → DE, DE → EN, and SV → EN and then manually evaluated whether the binomial adverbs were translated correctly. Tables 3, 4 and 5 contain the resulting MT accuracy. The numbers describe the percentages of sentences that had a correct translation of the binomial adverb. It does not mean that the complete sentences were translated correctly.

The first impression is that binomial adverbs are handled surprisingly well by Google Translate. For example, the German *ab und zu* is not only translated correctly but also with some variation into English as *occasional, sometimes, from time to time* which are all good translations. See example 1 which features a correct translation of the binomial adverb but also a number of tricky word reorderings which result in an excellent rendering of the meaning in English.

(1) DE: Somit riskieren wir **ab und zu**, im Sande steckenzubleiben.
Google EN: Thus we **sometimes** risk getting stuck in the sand.

The most striking problem in Google's machine translation of the binomial adverbs are omissions. The adverb is sometimes dropped in the translation as in example 2. This is an irritating finding, in particular since the generated target language sentence is fluent and grammatically correct. It looks good at first sight but misses an important aspect (expressed by the multi-word adverb) from the input sentence.

(2) DE: Ein dritter zeigt **ab und zu** Dias, die er selber in der Umgebung der Hütte gemacht hat.
Google EN: A third shows slides, which he himself has made in the vicinity of the hut.

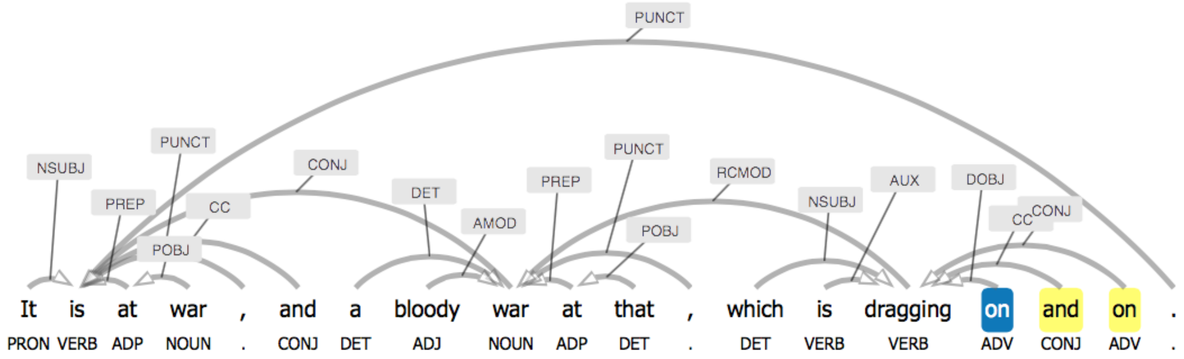

PUNCT
NSUBJ
PUNCT
PREP
CC
POBJ
CONJ
DET
AMOD
PREP
PUNCT
POBJ
RCMOD
NSUBJ
AUX
DOBJ
CC
CONJ
It is at war , and a bloody war at that , which is dragging on and on .
PRON VERB ADP NOUN . CONJ DET ADJ NOUN ADP DET . DET VERB VERB ADV CONJ ADV .

Figure 3: English sentence from Europarl with multi-word adverb (*on and on*) correctly tagged but incorrectly parsed.

Such omissions account for the majority of translation errors with respect to binomial adverbs. There are occasional wrong translations but to a much lesser extent. In example 3 we see that *on and off* is erroneously translated into German. It should have been *ab und zu, mit Unterbrechungen*.

(3) EN: The rains continued, **on and off** until mid April, unusually late for Jordan.
Google DE: Der Regen fuhr fort, **ab und ab** bis Mitte April, ungewöhnlich spät für Jordanien.

The translation of the English adverb *on and off* is difficult since sometimes it can have its literal meaning (*a torch flickered on and off*) whereas in other cases only the idiomatic translation is correct. This may explain its low MT accuracy.

For Swedish we checked *till och med* because it is so frequent and it also can serve two purposes. With that in mind we conclude that a 94.3% translation accuracy is good. In addition, we randomly extracted 10 sentences where *till och med* is an adverb and 10 sentences where the sequence is a conjoined preposition. Interestingly, the MT system translated *till och med* correctly in all 20 test sentences. The ones with the adverb reading were all translated with the English word *even* whereas the preposition cases were translated with *to / until / through*.

5 Conclusion

We have shown how to narrow down the search for binomial adverbs, a special type of multi-word expressions. We used the irreversibility score and a mutual information score to find cases that are top candidates for idiomatic usage.

We subsequently selected 7 such binomial adverbs from English, German and Swedish each and evaluated them in PoS tagging, dependency parsing and machine translation. The results are mixed in that PoS tagging and parsing works very good for some and badly for others. If we consider that down-stream applications rely on the parsing results, our study pin-points the need to handle such binomial adverbs with more care.

Statistical and Neural Machine Translation do not rely on parsing, and we therefore evaluated the binomial adverbs separately with Google Translate. We observed that frequent binomial adverbs like *by and large, first and foremost, over and over* in English, or *nach und nach, nach wie vor* in German, or *helt och hållet, till och med* in Swedish are translated well but not perfectly. The biggest problem is that Google Translate sometimes omits the binomial adverb which can be detrimental for the understanding of the sentence in the target language.

There is currently no repository of English, German or Swedish multi-word adverbs as in French (Laporte and Voyatzi, 2008) and some other languages. Our work would like to contribute to compiling such repositories.

Acknowledgments

This research was supported by the Swiss National Science Foundation under grant 105215_146781 for "SPARCLING: Large Scale PARallel Corpora for LINGuistic Investigation" (2013-2017). Thanks to Christof Bless for his tool to visualize the dependency trees. Thanks also to the anonymous reviewers for helpful comments on an earlier version of this paper.

References

Gerhard Bendz. 1965. *Ordpar*. P. A. Nordstedt & Söners Förlag, Stockholm.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multiword expressions for statistical machine translation. In *Proc. of LREC*. Istanbul.

Mathieu Constant and Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin.

Eric Laporte and Stavroula Voyatzi. 2008. An electronic dictionary of French multiword adverbs. In *Proc. of LREC*. Marrakech, Morocco.

Lukas Michelbacher, Stefan Evert, and Hinrich Schütze. 2007. Asymmetric association measures. In *Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP)*.

Sandra Mollin. 2014. *The (Ir)reversibility of English Binomials. Corpus, constraints, developments*, volume 64 of *Studies in Corpus Linguistics*. John Benjamins.

Gereon Müller. 1997. Beschränkungen für Binomialbildungen im Deutschen. *Zeitschrift für Sprachwissenschaft* 16(1):25–51.

Alexis Nasr, Carlos Ramisch, José Deulofeu, and André Valli. 2015. Joint dependency parsing and multiword expression tokenization. In *Proceedings of ACL*. Beijing, pages 1116–1126.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*. Theory and Applications of Natural Language Processing. Springer.

Liling Tan and Santanu Pal. 2014. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore.

Martin Volk, Simon Clematide, Johannes Graën, and Phillip Ströbel. 2016. Bi-particle adverbs, PoS-tagging and the recognition of German separable prefix verbs. In *Proceedings of KONVENS*. Bochum.

Dominic Widdows and Beate Dorow. 2005. Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of the ACL-SIGLEX 2005 Workshop on Deep Lexical Acquisition*. Ann Arbor, Michigan.

Chinese Separable Words in SMT

Gongbo Tang, Fabienne Cap, and Joakim Nivre
Department of Linguistics and Philology
Uppsala University
firstname.lastname@lingfil.uu.se

Abstract

The translation of Chinese separable words is a challenge in Chinese-English machine translation. In this paper, we propose a simple yet effective method that combines the two Chinese characters in separable words together whenever these two characters do not occur consecutively. Our experimental results show that our method can improve the translation quality of Chinese separable words significantly. We achieved improvements of 2.79 BLEU points and 17% accuracy (manual evaluation) on translating sentences with separable words.

1 Introduction

The correct translation of multi-word expressions (MWEs) such as compound words, phrases, collocations, and idioms, is very important for the performance of machine translation. It is easy for a statistical machine translation (SMT) system to translate a single word, but it is much harder to translate MWEs. The meaning of an MWE often differs from the combination of the meanings of its componont words. For example: 做文章 can be segmented as 做_文章.[1] 做 and 文章 can be translated into "make" or "do", and "article", respectively. However, the correct translation of 做文章 is "do some tricks" or "do something". We can see that the meanings are totally different.

Translating from Chinese into English is harder than translating from English into Chinese, and even much harder than translating other European languages into English. In addition to lexical and structural ambiguities that are common for many source languages, Chinese also faces the following challenges (Liu and Yu, 1998):

- Chinese word segmentation: Chinese has no word boundary markers. Although the accuracy of segmentation can be higher than 97% (Zhang et al., 2016), the error rate will be increased during translation which affects the translation quality severely.

- Chinese does not have tense markers: we have to use some function words to help us translating Chinese verbs into English.

- Chinese sentence structure: there is a big difference between Chinese sentence structure and English sentence structure.

Chinese separable words are phrasal verbs and they can be separated into two parts, a verb and a noun. Hence they can be viewed as MWEs. Due to their special characteristics, they are an interesting phenomenon to study and we are not aware of any previous work that has investigated them in the context of SMT. Most of the separable words have two characters. For simplicity, we will use capital letters to represent Chinese characters. Here is a separable word: "AB". These two characters "A" and "B" can be consecutive, "AB". However, they can also get separated by several other characters, like "ACDEB", and the number of these characters can be variable. Separable words can even have a reverse order, like "BDCA". In addition, there are some mixed forms like "ACAB", "AAB". Actually, there is no agreed-upon definition of separable words yet. One widely-used definition is that separable words are viewed as a word when they are consecutive, and are viewed as a phrase when they are separated (Zhang, 1957; Zhu, 1984). We will call the inconsecutive ones **separated separable words**, and the consecutive ones **unseparated separable words**. There are 3,701 words annotated as separable words in *XianDaiHanYuCiDian* (Institute of Linguistics CASS, 2002), which is

[1] In this paper, "_" represents a blank space.

Input	Separated separable word	Unseparated separable word
	请了几天假	请假了几天
Gloss	please, already, several, day, holiday	please, holiday, already, several, day
Translation	Ask for a few days off	
Google (Google, 2017)	Please take a few days off.	Leave for a few days.
Baidu (Baidu, 2017)	Please take a few days off.	A few days off.
Bing (Bing, 2017)	Please have a few days off.	Leave for a few days.
Sogou (Sogou, 2017)	A few days, please.	Leave for a few days.
Youdao (Youdao, 2017)	Have a few days off.	Ask for leave for a few days.

Table 1: Examples for translation outputs of freely available translation software.

the most authoritative Chinese dictionary. Nevertheless, there are many other words that have word formations similar to that of separable words.

There has been some work on translating Chinese separable words (Wang and Li, 1999; Shi, 2002), which got good results when translating a certain subset of separable words. However, they are using instance-based or rule-based methods which are not practical when dealing with a large amount of data.

If characters of separable words are not consecutive, they are more likely to be viewed as two independent characters, and translated incorrectly. Table 1 shows an example. Thus, our general idea is to move the parts of the separable words closer together in order to enhance the probability that they then will be learned as a phrase in the SMT model. Our experimental results show that our model achieves a considerable improvement on translating both separable words and the entire sentences.

The main contribution of this paper is that we verified that combining characters in **separated separable words** together can improve the translation of separable words. Moreover, the translation quality of sentences containing separable words is also improved considerably.

2 Related Work

2.1 MWEs in MT

Many researchers have proposed different methods to improve the performance of translating MWEs using MT.

Ma et al. (2007) packed several consecutive words together when those words were more likely to correspond to a single word in the opposite language. In this way, the number of 1-to-n alignments is reduced, and thus makes the task of alignment both easier and more natural. Our general idea is very similar to this method. In our experiments, we move the parts of the separable words closer together in order to enhance the probability that they will be learned as a phrase in the SMT model.

Ren et al. (2009) proposed an LLR-based hierarchical reducing algorithm to extract MWEs in source sentences, and then used GIZA++ to generate candidate translations of MWEs. Instead of using automatically extracted MWEs to filter the phrase table, they used MWEs as an additional resource and retrained the model. Their experiments on the Chinese traditional medicine domain and the chemical industry domain show that this method is feasible.

In addition to using a method based on word alignment to extract MWEs, Bai et al. (2009) used normalized frequency to generate possible translations and select the top-n candidates by Dice coefficient. They adopted those extracted translations to rewrite the test set, and could show a significant improvement in translation quality.

Due to the relatively free word order of German, the identification of support-verb construction is a challenge. Cap et al. (2015) added a special markup to the verbs occurring in a support-verb constructions, and gained a more correct verb translation.

Since MWEs are difficult to translate, why not adapt source sentences to make them easier to translate? Ullman and Nivre (2014) paraphrased compound nouns using prepositional phrases and verb phrases. The experimental results indicated a slight improvement in translation of paraphrased compound nouns, with a minor loss in overall BLEU score.

In addition to adapting the source language, there is also work on generating compound words for target languages. Matthews et al. (2016) first built a classifier to identify spans of input text that can be translated into a single compound word in the target language. Then they generated a pool of possible compounds which were added to the translation model as synthetic phrase translations. The results showed that this method is effective.

2.2 Chinese Separable Words

There has been a wide range of previous work on the identification of Chinese separable words. However, there is only little work on translating separable words.

Identification. For the identification of separable words, there are two mainstream methods.

The first one identifies separable words based on hand-crafted rules (Fu, 1999; Shi, 2002). For example, Fu (1999) used the rules of Chinese characters in a dictionary. If a character and its following characters match a rule of being a separable word, they will be identified as a separable word.

The second method is based on statistics. Zang and Xun (2017) used a large corpus to extract identification rules for 20 separable words. Their result showed that their method can be well applied to another 120 frequently occurring separable words. Wang and Li (1999) identified separable words with the constraint of bilingual data. If there exists a translation of the verbs in separable word candidates in the bilingual data, this candidate will be identified as a separable word.

Translation. Shi (2002) proposed a method based on rules and instances. He classified separable words into four groups by syntax structure, and translated them by instances. In addition, he also created different rules to translate **separated** and **unseparated** separable words.

Since word segmentation is the first step of translating Chinese to English, Wang and Li (1999) found that some words are segmented during segmentation, but these words have to be assembled together during translation. Wang and Li (1999) treated all these words as separable words during translation. They summarized 36 rules for separable words, and designed two detailed rules for translating these separable words.

3 Methodology

3.1 Identification

One of the most challenging problems for the identification of separable words is that the characters that constitute separable words are extremely productive. For instance, 看得到 can be segmented as 看_得_到 (“look”, “get”, “arrive”), and the correct translation of 看得到 is “can see”. However, 得到 (“get”) is a separable word. If 看得到 is identified as containing a separable word, it will be translated into “look get” which is completely wrong. Some of these problems can be solved during segmentation. For example, if 看得到 is segmented as 看得到, there will be no such problem. In our experiment, we used a fine-grained segmentation model, which is the default model of Zpar (Zhang and Clark, 2010).

We used the 140 separable words mentioned in Zang and Xun (2017) and 得到 which has a lot of instances. As the existing method cannot applied directly to our data set, we used a two-phase method with a high accuracy to identify separable words. First, we identified sentences with separable words by some simple rules. For **unseparated** separable words “AB”, if they exist as “_AB_” or “_A_B_”, they will be identified. For **separated** separable words “AB”, we use the regular expression “_A_.{1,5}_B_” to match each clause. Then, we manually filter the sentences identified by the first phase in training set. In contrast, for the identified sentences in development set and test set, they are only identified by rules without manual filtering.

3.2 Preprocessing

We modified all the separable words identified in our training, development, and test sets. Our general idea is to move the parts of the separable words closer together in order to enhance the probability that they then will be learned as a phrase in the SMT model. For **unseparated** separable words, they do not need to be modified if they are not segmented (“AB”), but they will be modified to “AB” if they are segmented (“A_B”). For **separated** separable words, the second character of the separable word (“B”) will be moved next to the first character (“A”). (“A_CD_E_B” → “AB_CD_E”)

	Sentence pairs	Separable word types	Separable word tokens
Training set	8995712	116	13339
Development set	1003	6	7
Test set	1006	8	8
New training set	8995662	116	13289
New test set	50	44	50

Table 2: Chinese-English corpus statistics.

4 Experiments

In the following, we report on two experiments. For the first experiment, we used all the parallel data to train the baseline model and our model, and evaluated the models on a standard test set(CWMT 2009). Since separable words are very sparse in the CWMT 2009 test set, the results of our models are inconclusive for separable word translation. We thus designed the second experiment in which we used an enriched test set. In contrast to the first experiment, the enriched test set is extracted from the training set, which is filtered manually. All the sentences in the test set contain a separable word token. At the same time, these extracted sentences are removed from the training set. Then we retained the models with the original settings. Both experiments used the same development set.

4.1 Data

We use data from the WMT 2017 shared task on Chinese-English. [2] It consists of roughly 8.9 million sentence pairs for training, 1,003 sentences for development and 1,006 sentences for testing. More detailed statistics are shown in Table 2.

The English language model is trained on the Xinhua News dataset from the Gigaword fifth edition. This training set is also part of the WMT 2017 shared task specified monolingual data. It consists of roughly 2 billion tokens.

4.2 Preprocessing

As usual, all data needs to be tokenized before SMT model training. We then true-cased the English data according to WMT standards. As for the Chinese data, note that tokenization is equivalent to word segmentation, which is a challenging task. We used the default segmentation model provided by Zpar (Zhang and Clark, 2010), which is one of the most accurate available models for Chinese. The resulting segmentation is very fine-grained, which is good for identifying separable words, may simplify the translation process, and reduces OOV words (because of generating much less word types). Potential over-segmentation is not problematic as Moses can easily recover from it due to the fact that it will then simply store sequences as phrases in the model.

Although there exist more than 3,560 two-character separable words in Chinese, we decided to just use the most frequent 141 separable words, in order to alleviate data sparsity. After filtering by rules and manual filtering as described in Section 3, there are only 116 different separable word types. In the first experiment, we found separable words in 13,339 sentences of the training set, in 7 sentences of the development set, and in 8 sentences of the test set. Note that we have much more separable word types and tokens in the enriched test set compared to the standard test set. The detailed statistics are shown in Table 2.

4.3 Experimental Settings

We used the default settings of Moses (Koehn et al., 2007) for training both a baseline model and our modified model with re-arranged separable words. We use Minimum Error Rate Training for tuning. On the target side, we used KenLM (Heafield, 2011) to train a 5-gram language model for English.

5 Results

5.1 Standard Test Set

For the automatic evaluation, we follow Matthews et al. (2016) who compared the quality of translations on all sentences vs. only on sentences that contained the phenomenon under investigation (in their case: compounds). In our case, we let both translation models translate the entire sentences with separable words and the phrases containing separable words to see whether our method can

[2] http://www.statmt.org/wmt17/translation-task.html

improve the performance of translating separable words. The BLEU scores (Papineni et al., 2002) of both models are given in Table 3. As the score of the 4-gram calculation is 0 when evaluating the translation of phrases with separable words, the final BLEU score is 0, we thus omit it. [3]

From Table 3, we can see that there is nearly no difference between these two models. In addition, the translation qualities of phrases with separable words are also slightly different.

	CWMT2009	Sentences with separable words
Baseline	23.97	24.12
New	23.69	24.18

Table 3: BLEU scores of both models in test set.

5.2 Enriched Test Set

In our previous experiment, we have found that only 8 sentences of the standard test set contain separable words. We have thus decided to re-run the experiment, but this time use an enriched test set with more instances of separable words, in order to better measure the effect of our processing on translation quality.

We extracted 50 sentences with separable word instances from the training set as a new test set, and retrained these two models. We removed these sentences from the training set before training our models.

We first computed the BLEU scores of both translation outputs. The BLEU scores for the enriched test set are given in Table 4. We can see a large difference between the Baseline and our result, which is larger than that in the previous experiments. Our model achieved 2.79 BLEU points improvement on the enriched test set, which is significantly better.

	Baseline	New
BLEU	12.85	15.64

Table 4: BLEU scores of both models in the enriched test set.

5.3 Manual Evaluation

In addition to using BLEU scores for evaluation, we also manually evaluated the translations of the enriched test set, both the translations of the separable words and the entire sentences as a whole.

The evaluation was set up as follows: the translations of the baseline model and our model are not marked for the evaluators. Two Chinese native speakers score 0 or 1 for the translation of separable words and the entire sentence. 0 means that the translation does not have the correct information, and 1 means that the translation has the basic original meaning. Table 6 shows some evaluation examples. We then average the scores of the two evaluators, and get the final accuracy result given in Table 5.

	Baseline	New
Separable word	49%	60%
Sentence	33%	50%

Table 5: Translation accuracy of separable words and sentences in the enriched test set, evaluated by two Chinese native speakers.

From Table 5, we can see that not only the accuracy of translating separable words has distinctly improved, but also the accuracy of translating the entire sentence has improved considerably. This result is an additional indicator that our method can improve the translation quality of separable words, and the translation quality of the entire sentences improved as well.

5.4 Analysis

After looking over the translations, we found that there are still many OOV words, especially in the standard test set. In addition to named entities, such as 凉风垭, there are also many words that were not translated due to segmentation errors. For example, 踩油门 ("step on the gas") and 日报道 ("reported on day") are two segmentation results. If they were segmented as 踩_油门 and 日_报道 correctly, all of them would have been translated. Since all the tokens in the enriched test set are frequent in Chinese, there is no OOV problem in the enriched test set.

Both automatic and manual evaluation have shown that the translation outputs of our model are better than that of the baseline model. As we mentioned before (Section 1), it is more likely to get

[3] Note that this is due to the small data set. Originally, BLEU has been designed to report the translation quality on a whole text, not only a few sentences or phrases.

Instances		Separable word	Sentence
Source	她是个失了业的教师.		
Reference	she is an out-of-work teacher.		
Baseline	she is a lost of teachers .	0	0
Our system	she is an unemployed teacher.	1	1
Source	她好像开过刀。		
Reference	Oh, looks like she had an operation.		
Baseline	she seemed to be a knife.	0	0
Our system	like she had an operation on it.	1	1
Source	有什么要说的，当我面说。		
Reference	if you have something to say to me, say it.		
Baseline	have something to say, when I say.	0	0
Our system	have something to say to me.	1	1
Source	我猜你参过军		
Reference	and you've seen service, I presume.		
Baseline	I guess you have been in the army.	1	1
Our system	I guess you to join the army.	0	0

Table 6: Comparison of translation outputs and manual scores on separable words and sentences. Separable words and translations of MWEs with separable words are underlined.

a bad translation of separable words when the two characters are separated. For example, in Table 6, for the first instance, 失业 ("out-of-work" or "unemployed") is the separable word. The baseline model translated 失 ("lost") into "lost" independently, and 业 ("work") is not even translated. But the new model can translate 失业 correctly because the source is modified to 失业了. For the third instance, 当面 ("in front of" or "to" or "with") is the separable word. The baseline model translated 当 ("when") into "when" independently, and 面 is also not translated. But the new model can translate 当面 correctly because the source is modified to 当面我. Hence, modifying separable words can really improve the performance of translating separable words.

However, some words are translated incorrectly after modifying separable words, especially the words between the two characters of separable words. For the fourth instance in Table 6, 过 ("already") is the tense information of the verb 参 ("join"), when 军 is moved to the second place, 过 is not translated at all. Hence, the tense of translation is not correct. Furthermore, there are two more challenges. On the one hand, some separable words are ambiguous. For example, 冒险 can be translated into "risk" or "adventure". On the other hand, the extraction of separable words is difficult. Some words look like separable words, but they just have the same structure. For example 跳舞 ("dance") in 跳的舞 is not a separable word, 跳的舞 means "the dances that danced" rather than 跳舞的 which means the verb "dance".

From the translations in Table 6, we can see that combining the separated characters together can improve the performance of translating separable words. However, there are still some problems if we only combine those two characters together. How to modify separable words still needs to be further explored.

6 Conclusion

Chinese separable words have many different types. How to translate separable words is a challenge for machine translation from and to Chinese. In this paper, we proposed a simple yet effective statistical method to improve the translation of Chinese separable words. Our results show that combining two separately occurring characters of separable words can improve the translation of separable words considerably. Moreover, the translation quality of the entire sentence also gets a considerable improvement.

In the future, we will explore more specific methods when translating different kinds of separable words. In addition, there are still challenges in how to define and identify separable words that will be explored further.

Acknowledgments

We acknowledge the computational resources provided by CSC in Helsinki and Sigma2 in Oslo through NeIC-NLPL (www.nlpl.eu). Gongbo Tang is supported by China Scholarship Council (No. 201607110016). Fabienne Cap has recieved funding from a VINNMER Marie Curie Incoming Grant within VINNOVAs Mobility for growth programme.

References

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*. Association for Computational Linguistics, Singapore, pages 478–486.

Baidu. 2017. Baidu translator. http://fanyi.baidu.com/. Accessed March 2, 2017.

Bing. 2017. Bing translator. http://www.bing.com/translator/. Accessed March 2, 2017.

Fabienne Cap, Manju Nirmal, Marion Weller, and Sabine Schulte im Walde. 2015. How to account for idiomatic German support verb constructions in statistical machine translation. In *Proceedings of the 11th Workshop on Multiword Expressions (MWE)*. Association for Computational Linguistics, Denver, Colorado, pages 19–28.

Aiping Fu. 1999. Chinese sentence tokenization in a Chinese-English MT system [in Chinese]. *Journal of Chinese Information Processing* 13(5):8–14.

Google. 2017. Google translator. https://translate.google.com/. Accessed March 2, 2017.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, pages 187–197.

Institute of Linguistics CASS. 2002. *Modern Chinese Dictionary [in Chinese]*. The Commercial Press.

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 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, Prague, Czech Republic, pages 177–180.

Qun Liu and Shiwen Yu. 1998. The difficulties in Chinese-English machine translation [in Chinese]. In *Proceeding of Chinese Information Processing International Conference*. Tsinghua University Press, Beijing, pages 507–514.

Yanjun Ma, Nicolas Stroppa, and Andy Way. 2007. Bootstrapping word alignment via word packing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 304–311.

Austin Matthews, Eva Schlinger, Alon Lavie, and Chris Dyer. 2016. Synthesizing compound words for machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1085–1094.

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*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions (MWE): Identification, Interpretation, Disambiguation and Applications*. Association for Computational Linguistics, Singapore, pages 47–54.

Xiaodong Shi. 2002. The processing of separable word in Chinese-English translation [in Chinese]. In *Proceedings of National Symposium on Machine Translation 2002*. Publishing House of Electronics Industry, Beijing, pages 68–76.

Sogou. 2017. Sogou translator. http://fanyi.sogou.com/. Accessed March 2, 2017.

Edvin Ullman and Joakim Nivre. 2014. Paraphrasing swedish compound nouns in machine translation. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*. Association for Computational Linguistics, Gothenburg, Sweden, pages 99–103.

Haifeng Wang and Sheng Li. 1999. The stradegy of processing Chinese separable word in Chinese-English machine translation [in Chinese]. *Journal of the China Society for Scientific and Technical Information* 18(4):303–307.

Youdao. 2017. Youdao translator. http://fanyi.youdao.com/. Accessed March 2, 2017.

Jiaojiao Zang and Endong Xun. 2017. Automatic recognition of separable words based on bcc [in Chinese]. *Journal of Chinese Information Processing* 31(1):75–83.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Transition-based neural word segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 421–431.

Shoukang Zhang. 1957. A brief view of Chinese word formation [in Chinese]. *Studies of the Chinese Language* 6:3–9.

Yue Zhang and Stephen Clark. 2010. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA, pages 843–852.

Dexi Zhu. 1984. *Chinese grammar textbook [in Chinese]*. The Commercial Press.

Out of the tailor or still in the woods?[*] An empirical study of MWEs in MT

Fabienne Cap
Department of Linguistics and Philology
Uppsala University
fabienne.cap@lingfil.uu.se

Abstract

We present a test set for MT evaluation from German to English in which verbal MWEs have been manually annotated. This provides us with useful insights about the quantity of MWEs and their distribution in the test set. In addition to that, we manually aligned the annotated German test set to its reference translation and to the output of a state-of-the-art NMT system. We then compared the translations with each other in order to get an estimate for MWE translation quality. We found that the vast majority of MWEs across different categories have been correctly translated but that most of these translations are not captured by BLEU scores.

1 Introduction

Multiword expressions (MWE) consist of several words forming a unit with particular properties. One aspect of MWEs is that often the meaning of the whole expression cannot be derived from the meaning of its constituent words. Consider for example, the German MWE *aus dem Schneider sein* for which a word-by-word translation would result in "to be out of the tailor", while one of its correct translations is "to be off the hook". This and other characteristics of MWEs are a challenge for many NLP applications in general and for machine translation (MT) in particular.

While this fact is widely acknowledged (Sag et al., 2002; Villavicencio et al., 2005), and there have been several works to deal with MWEs in MT in the past (Carpuat and Diab, 2010; Cholakov and Kordoni, 2014), we are not aware of any previous work that performed a quantitative and qualitative analysis of MWEs in an MT test set. Not knowing the quantity of MWEs occuring in the test sets of the annual shared task of machine translation (WMT) makes it hard to estimate what effect a better treatment of MWEs would result in – in terms of translation quality.

In the present work, we attempt to close this gap by performing an exhaustive empirical study on one of the test sets that are commonly used to evaluate MT performance. We focus on verbal MWEs of German, which are particularly challenging due to the relatively free word order of German, which may result in many intervening words between the constituent words of the MWE. We manually annotate all verbal MWEs in the source language part of the German to English translation task, and we do so in accordance with freely available annotation guidelines. The resulting resource[1] does not only give us a quantitative impression on MWEs in the test set. It may in the future be used as an upper bound to test approaches to identify and better translate MWEs in MT.

Besides the MWE annotations, we also manually align the found instances to their counterparts in the English reference translation and to their counterparts in the output of a state-of-the-art NMT system. By classifying how MWEs are translated, we can gain some superficial insights into the performance of NMT on MWEs. These may in the future be useful when designing MWE-aware NMT models.

The remainder of this paper is structured as follows: In Section 2 we review previous work on MWE annotation and MWE treatment for MT. In Section 3 we present our methodology and give details e.g. about the guidelines we used. Then,

[*] "Out of the tailor" is a literal translation we found for the German MWE *aus dem Schneider sein)* ("to be out of the woods", "to be off the hook".

[1] All annotations performed will be made available in conjunction with this paper.

in Section 4 we present our results before we conclude and give an outlook for future work in Section 5

2 Related Work

Our work is to be settled at the edge of annotation efforts and MWE treatment in MT as it serves both purposes. In the following, we give a short review of both areas.

MWE annotation There are two kinds of MWE annotations to be distinguished: type-based and token-based annotation. Our approach belongs to the latter one, as we are annotating MWEs in the running text in which they occur. For English, there have been several such token-based MWE annotation efforts in the past. Cook et al. (2008) introduced a dataset containing literal and idiomatic usages of verb-noun pairs. Similarly, Tu and Roth (2011) presented annotations for light verb constructions, which they later extended to cover verb particle constructions (Tu and Roth, 2012).

An analysis of already annotated MWEs in treebanks is given by Rosén et al. (2015). They found considerable heterogeneity both in terms of MWE types that have been annotated and in terms of how they have been annotated.

More recently, there has been a huge token-based annotation effort for verbal MWEs in the framework of a shared task on MWE identification (Savary et al., 2017). They created guidelines and then annotated MWEs in a collection comprising 5 million words in 18 languages. In the present work, we will base our annotation on these guidelines. This makes it possible to train an MWE extraction system on the PARSEME data and then apply and evaluate it on our annotated test set, before running MT and extrinsically evaluating the impact of MWEs on translation quality.

MWEs in MT In the past, there have been several approaches to deal with different kinds of MWEs in statistical machine translation (SMT), treating e.g. particle verbs (Carpuat and Diab, 2010; Cholakov and Kordoni, 2014) or light-verb constructions (Cap et al., 2015).

However, similar to many other research areas within NLP, neural models have emerged and significantly improved MT performance in recent years (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2015; Cho et al., 2014; Bojar et al., 2016). Sennrich et al. (2016b) presented an NMT system that is able to deal with unseen German nominal compounds, which, despite the fact that they are written in one word, are often viewed as MWEs. More recently Sennrich (2017) investigated the translations of some linguistic phenomena (e.g. long distance agreements) in the context of NMT, but verbal MWEs were not included.

3 Methodology

In our study, we first annnotated MWEs on the source side of an MT test set. This gives us some intuition about the frequency and distribution of MWEs in such a test set. Then, we manually align the annotated source side to the reference translation of the same test set. This is important as most MT systems are still evaluated based on the BLEU metric (Papineni et al., 2002), despite its known deficiencies in terms of not adequately capturing morphological variants or synonyms. As BLEU is based on exact n-gram matching, we can thus only get credit for correct MWE translations if they occur in their exact same shape in the reference translation.

Finally, we manually align the source side of the test set to the NMT output obtained by one of the current state-of-the-art NMT systems. We do this in order to gain some insights in how well MWEs are captured by an NMT system which is well designed overall, but does not explicitly pay attention to MWE treatment. The findings gained from this evaluation will then show us which kinds of phenomena we have to focus on if we want to improve MWE translations in state-of-the-art NMT systems.

3.1 Dataset

We base our empirical study on a standard test set commonly used to evaluate machine translation systems. Such test sets are freely available and annually distributed in the context of the WMT shared tasks on machine translation. We picked the testset from 2016[2]. It contains 2,999 lines of parallel text in German together with an English reference translation.

While we will primarily annotate MWEs in the German part of this test set, we also make use of the English reference translation at a later stage, when we manually align the annotated MWEs to

[2] www.statmt.org/wmt16

English. In order to gain insights into the performance of a state-of-the-art NMT system without particular MWE treatment, we also needed an NMT system output for the same test set.

For this, we chose the output from the best performing system for the translation direction from German to English (Sennrich et al., 2016a) in the annual shared task for machine translation in 2016 (Bojar et al., 2016) . The system is an ensemble of NMT systems. It can deal with unknown words by making use of subword processing and uses back-translations from monolingual corpora in order to increase the training data size. Note that the complete output files for all systems that have participated in the annual shared tasks on machine translation are freely available[3].

3.2 Annotation

For the annotation of MWEs we relied on the annotation guidelines[4] that have been created and released in the framework of the PARSEME project[5] and its associated shared task on the identification of verbal MWEs (Savary et al., 2017).

The annotation guidelines are designed to be applicable to a wide range of different languages, of which German is one. They contain criteria to classify verbal MWEs (VMWEs) into five different categories: idioms, inherently reflexive verbs, light verb constructions, particle verbs and other. Following the recommendations given in the annotation guidelines we performed a two-step annotation: first, potential MWE candidates were identified, then they were classified using decision trees acompagnied by illustrative examples for each step. Overlapping MWEs are also annotated and will be counted twice in the evaluation.

The annotation was performed by a native speaker of German and linguist. We are aware of the fact that annotations based on only one person are less reliable than annotations based on multiple annotators. In the future, we might re-annotate this set with more annotators and then be able to report on inter-annotator-agreements. In the following, we briefly summarise the criteria for the different categories as they occur in the annotation guidelines mentioned above:

type	**token**
ID	417
IReflV	75
LVC	123
VPC	710
total	1,325

Table 1: Distribution of investigated MWE types.

Idioms (ID) have at least two lexicalised components, consisting of a head verb and at least one of its arguments, which can be of different type. Examples for German include: *er* ***hat*** *einen* ***Lauf*** (”he is on a roll”), ***es gibt*** (”there are”), *das* ***stellt*** *alles* ***in den Schatten*** (”to overshadow”, lit: “that puts everything in the shadow”).

Inherently Reflexive Verbs (IReflV) do either never occur without their reflexive clitic, or its sense or subcategorisation frame differ from the non-reflexive version of the same verb. Examples for German include: ***sich kümmern um*** (”take care of”), ***sich befinden*** (”to be located”).

Light Verb Constructions (LVC) consist of a ”light” verb which contributes little or no meaning to the whole expression and a noun which is predicative and often refers to an event. Examples for German include: ***Respekt zollen*** (”to respect”), *einen* ***Beitrag leisten*** (”to make a contribution”).

Verb-Particle Constructions (VPC) consist of a head verb and a lexicalised particle which depends on that verb. Its meaning is often non-compositional: one indicator for this is that the meaning of the verb with and without particle differ considerably. Examples for German include ***vorwerfen*** (”to accuse”), ***anfangen*** (”to begin”).

Other (OTH) is a category to collect all MWEs that could not be assigned to any other of the above categories. This includes expressions with more than one syntactic head, e.g. “to drink and drive”.

We faced two main challenges during the annotation phase: first, it was not always straightforward to distinguish between light verb constructions and idioms and second, it was often difficult to determine the (non-)compositionality of German verb particle constructions, which are highly ambiguous.

The result of our annotation is given in Table 1. We can see that 1,325 verbal MWEs occur in the 2,999 sentences of the testset. Moreover, we found

[3] matrix.statmt.org

[4] http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext/

[5] PARSEME was a European Cost-action on Parsing and Multiword Expressions from 2013–2017. More information on the project can be obtained via https://typo.uni-konstanz.de/parseme/

	Type	**German input**	**Reference translation**
correct translation	VPC	nahm teil	attended
	VPC	nahm teil	took part
correct but different	ID	viele Variablen **spielen** eine **Rolle** many variables **play** a **role**	many variables **are at play**
	ID	**gab es** Verbesserungen **there were** improvements	we **saw** improvements
not translated	ID	**bin** ich lange genug **im Geschäft** I **am** long enough **in the business** um realistisch zu sein	I **am** long enough to be realistic
	ID	**gibt es** für sie eine Mahlzeit bei **there is** a meal for you at	grab a meal at

Table 2: Examples to illustrate the categories used to asess the reference translation quality.

that there were 1,101 sentences which contained at least one verbal MWE. Verb-particle constructions represent the largest MWE group, followed by idioms, light verb constructions and inherently reflexive verbs. We found no instances of MWEs belonging to the category **OTH**. These numbers show that improving the translation of verb particle constructions and/or idioms will most probably lead to noticeable effects on translation quality measured using BLEU (Papineni et al., 2002). Note that BLEU scores are usually calculated by averaging over the whole test set. When improving the translation of a phenomenon that only rarely occurs the (positive or negative) effect will be much smaller than for a frequent phenomenon.

3.3 Word Alignment

In order to assess the translations of the MWEs that have been annotated in the test set, we performed a manual word alignment between both the German input and the English reference translation, as well as the German input and the NMT output. Note that if we had compared to a conventional SMT system, this step would not have been neccessary, as SMT systems can simply print the automatic word alignment links that were used to obtain the respective translations. The NMT system we relied on for the output features an attention-based mechanism. This mechanism is comparable with word alignments but not as exact as word alignments. Moreover, information on the attentions is not distributed in conjunction with the test set translations. Aiming for high accuracy in our present study, we thus decided to align the NMT output manually to the German input, assuming that the translational equivalents would be apparent enough. For convenience we used a simple java-based GUI to perform the word alignments, which significantly speeds up the manual alignment process.

3.4 Evaluation

After the manual annotation and alignments, we have to evaluate the quality of the translations we obtained in this way. As we are mainly interested in the translation of MWEs and not overal translation quality, we do not rely on BLEU scores, but instead, we perform manual evaluations for both alignments (source→reference and source→NMT output). Due to the different nature of the two translations, we designed two sets of evaluation categories. Note that not only the manually obtained translations were presented at evaluation time, but also both sentences (source and reference or NMT output) were given. This is to ensure that decisions can be taken in context.

3.4.1 Reference translations

The reference translations for the test sets provided by WMT have been produced by professional human translators. We thus assume that their translations are correct. There is thus no need to introduce a category for wrong translations. However, when faced with an MWE, the human translator often has different options at hand to translate the sentence: translating the MWE into a corresponding construction in the target language (not neccessarily an MWE as well) or reformulating the sentence in a way where the meaning of the MWE is still conveyed but using different lexemes or a different structure than one would expect based on the MWE. Finally, it might happen that an MWE is not translated at all, either due to a different construction in which it can be omit-

	Type	German input	Reference translation	NMT translation
identical with ref.	VPC	wuchs hier auf	grew up here	grew up here
close to reference	ID	in Auftrag geben	commissioning	commissioned
correct but different	LVC ID	machte eine Anmerkung haben Anspruch	made a reference have right	made a note is entitled
acceptable	ID VPC	sich ihre Meinung gebildet hindeuteten	made up their minds suggesting	formed their minds hinted
wrongly translated	ID LVC VPC	im Sturm erobern Entscheidung fallen dreht durch	to win decision made going berserk	*to conquer in the storm *decision fall *turns through

Table 3: Examples illustrating the categories used for comparison of the NMT output and the reference.

ted or it was erroneously omitted by the human translator. The respective category names for this evaluation are given below:

- **correct translation:** the MWE was translated correctly.
- **correct but different:** the meaning of the MWE was coveyed by using a different lexemes or a different structure and the alignment was thus not straightforward.
- **no translation:** the MWE and its content was omitted in the translation.

Illustrative examples for each of these categories are given in Table 2.

3.4.2 NMT system

In contrast to the reference translations, the NMT system output might contain inadequate or influent translations. Moreover, there are two possibilities of the translation being *correct*: either by being identical or close to the reference translation, or, if not, by looking at the source sentence and judging whether or not the translation is correct (even though it might completely differ from the reference translation). We thus need to introduce different categories for this evaluation. They are given below:

- **identical with ref.:** the MWE translation is 100% identical with the reference translations. Note that those are the only ones accounted for when calculating BLEU scores.
- **close to reference:** the MWE translation differs only slightly from the reference translation, e.g.through morphological variantion.
- **correct but different** the MWE is translated correctly, but different lexemes and/or a different structure than in the reference is used.
- **acceptable:** the MWE translation is not fluent, but the meaning is still conveyed.
- **wrongly translated:** the MWE has been translated wrongly. Most of these cases include false litteral translations of its parts
- **not translated:** the MWE has not been translated.

In Table 3 we give some illustrative examples for all of these evaluation categories and several different MWE types.

4 Results

After having annotated the verbal MWEs in the test set, manually aligned them and classified the found translations into evaluation categories, we finally counted them and present our results below. As previously, due to the different nature of the two alignments (to reference vs. to NMT output) we present them separately.

4.1 Reference Translations

This evaluation aimed at estimating how many of the MWEs occurring on the source side of the test set have actually been translated to the target language by a human professional translator. An overview of the evaluation of the reference translations is given in Table 4. We can see that the vast majority of the MWEs have been translated, and they have been so mostly into similar and identifiable constructions. There were no IReflVs and LVC left untranslated and only a few instances for idioms (4) and verb particle constructions (4).

4.2 NMT Translations

The purpose of the NMT evaluation is to see how well MWEs are covered by a state-of-the-art NMT system without particular MWE processing. Also,

	ID	IReflV	LVC	VPC	total
correct translation	378	66	105	635	1,184
correct but different	37	9	18	74	138
no translation	2	0	0	1	3
total	417	75	123	710	1,325

Table 4: Evaluation of reference translations.

	ID	IReflV	LVC	VPC	total
identical with ref.	120	25	34	245	424
close to reference	98	15	28	108	249
correct but different	68	18	36	179	301
acceptable	35	2	9	52	98
wrongly translated	92	11	16	99	218
not translated	4	4	0	27	35
total	417	75	123	710	1,325

Table 5: Evaluation of NMT translations.

this manual evaluation might reveal correct translations that have not been accounted for in the BLEU scores, because they are not exactly matching the reference.

As mentioned earlier, these results can only be seen as an approximation. Due to the fact that NMT does not rely on word alignments, we do not know which source words were translated into which target words. Our manual evaluation is simply an educated guess for how an alignment would have looked like.

The results are given in Table 5. Note that only the MWEs assigned to **identical with ref.** are captured by BLEU scores. Translations belonging to the categories of **close to reference**, **correct but different** and **acceptable** are not. From Table 5 we can see that this applies to most of the MWE translations. This means that the NMT system we evaluated performs even better (in terms of MWEs) than reflected in the BLEU scores. Only roughly a fourth of all idioms and slightly fewer verb particle constructions have been translated wrongly. For the other two MWE categories it is even less. If we were to adapt this or a similar NMT system more to the needs of MWEs in the future, this means that we have to be very careful not to decrease the translation quality that is already fairly high for most MWEs (even though not fully reflected by BLEU scores).

5 Conclusion and Future Work

We have presented an empirical study to assess the quantity and quality of German verbal MWEs in the context of MT. To do so, we have manually annotated a standard test set which is (and will be) commonly used to evaluate MT systems. We followed available annotation guidelines in order to make the annotation coherent with a large collection of previously annotated verbal MWEs (Savary et al., 2017). This way, systems for the identification and treatment of MWEs can be trained on the previous dataset and then evaluated on the test set. Moreover, the annotated test set may serve as an upper bound for future approaches to deal with MWEs in MT: instead of running both the identification of the MWEs and their translation in one go, different methods for treating MWEs in MT can be tested separately using the annotated test set as oracle input to the system.

When manually evaluating the output of a state-of-the-art NMT system against the reference translation, we found that many more MWEs have been directly translated by this system even though it does not explicitely model MWEs.

A straightforward future extension of our annotated test set would be to no longer be restricted to verbal MWEs, but also include nominal MWEs like compounds or adjective verb pairs. Our annotation revealed that the types of MWEs found in the test set are not evenly distributed in terms of quantity. Depending on future applications, it might be useful to create a new test set which is more balanced with respect to different MWE types.

Acknowledgements

This project has been funded by a VINNMER Marie Curie Incoming Grant within VINNOVAs Mobility for Growth programme. Thanks to the anonymous reviewers for their constructive comments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR'14: Proceedings of the International Conference on Learning Representations*.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation (wmt16). In *Proceedings of the First Conference on Machine Translation (WMT)*. volume 2, pages 131–198.

Fabienne Cap, Manju Nirmal, Marion Weller, and Sabine Schulte im Walde. 2015. How to Account for Idiomatic German Support Verb Constructions in Statistical Machine Translation. In *Proceedings of the 11th Workshop on Multiword Expressions (MWE) at NAACL*. pages 19–28.

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*. Association for Computational Linguistics, pages 242–245.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *SSST'14: Proceedings of the eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. pages 103–111.

Kostadin Cholakov and Valia Kordoni. 2014. Better statistical machine translation through linguistic treatment of phrasal verbs. In *EMNLP'14: Proceedings of the 2014 conference on Empirical Methods for Natural Language Processing*. pages 196–201.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*. pages 19–22.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP'13: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pages 1700–1709.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *ACL'02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. pages 311–318.

Victoria Rosén, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejcek, Agata Savary, Adam Przepiórkowski, Petya Osenova, and Verginica Barbu Mititelu. 2015. A survey of multiword expressions in treebanks. In *Proceedings of the 14th International Workshop on Treebanks & Linguistic Theories conference, Warsaw, Poland, December*. pages 179–193.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 1–15.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZade, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE) at EACL*. pages 31–47.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. In *EACL'17: Proceedings of teh 15th Conference of the European Chapter of the Association for Computational Linguistics*. pages 376–382.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. In *WMT'16: Proceedings of the First Conference on Machine Translation*. pages 371–376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *ACL'16: Proceedings of the Annual Meeting of the Association for Computational Linguistics*. pages 1715–1725.

Yuancheng Tu and Dan Roth. 2011. Learning English light verb constructions: contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. Association for Computational Linguistics, pages 31–39.

Yuancheng Tu and Dan Roth. 2012. Sorting out the most confusing English phrasal verbs. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pages 65–69.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: having a crack at a hard nut. *Computer Speech & Language* 19(4).

Corpus analysis of the Path elements of the verb *otići* / *oditi* in Croatian and Slovene

Goranka Blagus Bartolec and Ivana Matas Ivanković
Institute of Croatian Language and Linguistics
Republike Austrije 16, 10000 Zagreb, Croatia
{gblagus,imatas}@ihjj.hr

Abstract

The aim of this research is to compare verb-prepositional collostructs with the perfective prefixed verb of motion *otići* 'leave' in Croatian and *oditi* 'leave' in Slovene, both on the level of surface expression and on the level of meaning. First, based on Talmy's research (1985/2007), we compare the prefixes *od-* lit. 'from-' and prepositions in Croatian and Slovene that represent Path elements of motion as closed-class units together with the basic verb of motion. They denote the path or direction of a trajector's motion in relation to a landmark. Second, we describe the meaning potential of the Croatian and Slovene prefix *od-,* which forms a constituent element of the verb *otići* / *oditi* in both languages. We emphasize collostructs with repetitive prefixal-prepositional constituents – the Croatian *otići od* and the Slovene *oditi od* 'go from'. Then, by means of raw frequency in the hrWaC and slWaC corpora, we compare it with other prepositions used with the verb *otići* / *oditi*. Relying on these results, we analyze whether the prefix or the preposition is the key Path element determining the path or direction of the basic verb of motion.

1 Introduction

In this corpus-based research, we compare verb-prepositional collostructs[1] consisting of prefixed verbs and prepositions in Croatian and Slovene. The verbs of motion are often followed by a prepositional phrase denoting many spatial relations. On the formal level, this relation is frequently expressed with a homonymous prefix-prepositional pattern. Our goals were to (1) examine the relationship between prefix and preposition in such pairs based on Talmy's path analysis (1985/2007), (2) carry out a contrastive analysis of two cognate languages. We have chosen Croatian and Slovene as genetically related languages that belong to the western group of South Slavic languages, and we expect a high level of correspondence within the domains of both meaning and surface expression. We have chosen the Croatian and Slovene prefixed verb *otići* / *oditi* 'leave'[2] for two reasons: (1) the verb *otići* / *oditi* 'leave', as a verb complex of the root verb *ići* / *iti* 'go', has many meanings, and leaves room for various syntactic structures; (2) it often appears with prepositional phrases, including the homonymous preposition *od* 'from', with which it forms a repetitive pattern. The prefix *od-* lit. 'from-' is one of the most common prefixes with verbs of motion in both Croatian[3] and Slovene. In searches of corpora (the Croatian hrWaC v2.2 corpus[4] and the Slovene slWaC 2.1 corpus[5]), we first included Croatian and Slovene

[1] The terms *collostruct* and *collostruction* are used according to Stefanowitsch and Gries (2003:215).

[2] Hereinafter, the first example is in Croatian, the Slovene equivalent comes after the slash, and the English translation is provided in single quotation marks.

[3] *Ot-*, *oda-* and *o-* (Babić, 1986:484–485) are also allomorphs of the prefix *od-* in Croatian.

[4] HrWaC is a web corpus collected from the .hr top-level domain. Version 2.2 of the corpus contains almost 1.4 billion tokens and is annotated with lemma, morphosyntax, and dependency syntax layers. The compilation of version 1.0 of the corpus is described in Ljubešić and Erjavec (2011), while the 2.0 version is described in Ljubešić and Klubička (2014).

[5] The Slovene slWaC corpus is a web corpus collected from the .si top-level domain. Version 2.1 of the corpus contains almost 0.9 billion tokens and is annotated with lemma and morphosyntax layers. The compilation of version 1.0 of the

prefixed verbs with the prefix *od-* ‘from‘. In the CQL field, we entered the regular expression [word="od.*|ot.*"&tag="V.*"] in hrWaC, and [word="od.*"&tag="G.*"] in slWaC. The tagsets differ slightly between the two corpora, but these regular expressions will provide examples of verbs that start with *od-*. Since the corpora were collected in the same way, they can be compared. They differ in size, so we relied on the order of lemmas based on raw frequency.

Croatian verbs with prefix *od-*		**Slovene verbs with prefix *od-***	
otići ‘leave’	379,449	odločiti ‘decide’	275,191
odlučiti ‘decide’	372,269	odpreti ‘open’	132,547
održati ‘maintain’	228,870	odpraviti ‘ship’	125,774
otvoriti ‘open’	212,505	oditi ‘leave’	91,316
odgovarati ‘reply’	197,085	odgovoriti ‘reply’	84,881
odgovoriti ‘reply’	175,721	odločati ‘decide’	79,944
otkriti ‘discover’	171,563	odkriti ‘discover’	71,717
odnositi ‘carry’	154,400	odstraniti 'remove’	61,864
održavati ‘keep’	146,453	odpeljati ‘take’	59,974
odigrati ‘play’	112,795	odigrati ‘play’	52,169

Table 1: 10 most frequent Croatian and Slovene verbs with prefix *od-* ‘lit. from-’

We singled out the ten most frequent verbs in each language (Table 1). Among them, two Croatian verbs have the meaning of motion (*otići* ‘leave’ i *odnositi* ‘carry off’), while three Slovene verbs have the meaning of motion (*oditi* ‘leave’, *odpraviti* ‘ship’, and *odpeljati* ‘take away’). The search showed that the most frequent Croatian verb with the prefix *od-* is *otići*. The Slovene corpus singled out *odločiti* ‘decide’ as the most frequent verb, which is not a verb of motion. The verb *oditi* ‘leave’ is in fourth place according to frequency, but it is equivalent to the Croatian *otići*.

The rest of the work is structured as follows: first, some general remarks about Talmy’s description of Path are provided along with examples from Croatian and Slovene, followed by a description of the prefix *od-* and the preposition *od*, followed by the corpus analysis.

2 Path

The Croatian collostruct *otići od* and the equivalent Slovene collostruct *oditi od* ‘go from’ have the prototypical structure of a prefixed verb followed by a preposition homonymous to the prefix. This repetitive pattern is considered typical in Russian (Talmy, 1985/2007) and some other Slavic languages as well (Brala-Vukanović and Rubinić, 2011; Brala-Vukanović and Memišević, 2014^1/2014^2; Mitkovska and Bužarovska, 2012). Talmy (1985/2007) primarily associates this pattern with verbs of motion (e.g. in Croatian and Slovene: *ući u* / *vstopiti v* ‘get into’, *doputovati do* / *dopotovati do* ‘arrive at / travel to’), but the repetitive pattern appears with other prefixed verbs as well – generally, with all verbs connecting a trajector (TR) and a landmark (LM)[6]. A trajector can be correlated with a landmark in any type of process: e.g. *napisati (jednadžbu) na (ploču)* / *napisati (enačbo) na (tablo)* ‘write an (equation) on (the blackboard)’; *zavezati (uže) za (stup)* / *zavezati (vrv) za (steber)* ‘tie (a rope) to (a post)’. In these collostructions, the prefix and preposition define the path or direction of the action expressed by the verb. Describing this semantic relationship, Talmy (2007:70) uses the term Path (P) and defines it as follows: “The ‘Path’ (with a capital ‘P’) is the path followed or site occupied by the Figure object with respect to the Ground object.” Path refers to the kind of motion, i.e. how the trajector moves in relation to the landmark (e.g. entering, exiting, oncoming, distancing, descending, ascending etc.). In the collostructions *ući u kuću* / *vstopiti v hišo* ‘enter the house’, the prefix *u-* / *v-* ‘into’ and the preposition *u* / *v* ‘into’ are Path elements referring to oncoming or approaching a landmark, which is the goal of the motion (in the adlative sense). In the collostructions *odmaknuti se od prozora* / *odmakniti se od okna* ‘move away from the window’, the prefix *od-* / *od-* ‘from’ and the preposition *od* / *od* ‘from’ are Path elements referring to distancing or moving away from the landmark, which is the starting point and the source of motion (in the ablative sense). “Generally, the Path is expressed fully by the combination of a satellite and a preposition.” (Talmy 2007:141) The verb root (i.e the basic verb[7]), the trajector, and the landmark are open-class elements, whereas the prefix is a satellite – a closed-class element – “the grammatical category of any constituent other than a nominal complement that is in a sister relation to the verb root. It relates to the verb root as a dependent to a head.” (Talmy 2007: 139) In addition to the Path, a satellite can also

corpus is described in Ljubešić and Erjavec (2011), while the 2.0 version is described in Erjavec et al. (2015).

[6] We use Langacker’s (1987) terms *trajector* and *landmark*, which have been widely accepted in Croatian linguistics (Belaj and Tanacković Faletar, 2014; Brala-Vukanović and Rubinić, 2011). These terms are equivalent to Talmy’s terms *Figure* and *Ground* (1985/2007).

[7] The root verb of *otići* / *oditi* is *ići* / *iti* ‘go’.

express Ground, Patient, Manner, Cause, etc. Talmy (2007:144) primarily considers only those elements that are always attached to the verb (e.g. prefixes in Slavic languages such as *u-* in the Croatian *ući* and *v-* in the Slovene *vstopiti* 'get into') or forms that often overlap with the preposition in phrasal verbs in English (e.g. the *out* in *run out*) to be verb satellites. He does not consider independent prepositions that form prepositional phrases with a noun to be verb satellites, since the prepositional phrase can be omitted (we can say both *Ana je ušla* and *Ana je ušla u kuću* / *Ana je vstopila* and *Ana je vstopila v hišo* 'Ana came in' / 'Ana came into the house'). In the broader semantic context in which the collostruction (consisting of a prefixed verb of motion and a prepositional phrase) appears, the preposition together with the verb prefix is the key Path element that determines the direction of motion of the trajector relative to a landmark. Besides, even if it is not mentioned, the landmark is always implied. We focus on the Path-satellite model which Talmy (2007: 158) describes as "satellites determining the Figure–Ground precedence pattern of the verb". On the level of surface expression, this includes the prefix and the preposition as elements that determine the path of motion of the trajector relative to a landmark.

3 The semantic features of the prefix *od-* / *od-* and the preposition *od* / *od*

The perfective aspect in Croatian and Slovene is often lexicalized in the form of a prefix. According to Babić (1986:484–485), the Croatian prefix *od-* with verbs of motion has the meaning of (1) separation, distancing, moving away, disassembling, and (2) completion of the verb action, among others. According to SSKJ (1997), the Slovene prefix *od-* with a verb of motion refers to (1) the direction of action departing from a specific point, (2) departing, distancing from a specific place, or (3) the completion of an action. Therefore, in the Croatian verb *otići* and the Slovene verb *oditi* 'leave', the prefix *od-* can be defined as having the meaning of departing, distancing from somewhere and the meaning of perfective aspect. According to Talmy's motion-aspect formulae (2007:93), *otići* / *oditi* can be considered a 'move from' verb.

The prefix *od-* is complementary to the preposition *od* 'from' in the meaning of leaving, separating, departure, changing position, or moving from a position (RHJ, 2000; SSKJ, 1997). According to the cognitive grammar of the Croatian language (Belaj and Tanacković Faletar, 2014:309), the genitive with the preposition *od* in movement scenarios denotes the prototypical concept of the spatial distancing of a trajector from a landmark. A landmark in the genitive represents the most obvious concept of a source. In the concept of trajector's motion, a landmark in the genitive refers to spatial starting points. The meaning of ablocality (ablative locality) is typical of this preposition.

Od is a typical representative of 'from'-type prepositions (e.g. *s* / *s* 'from', *iz* / *iz* 'out of', etc). Based on these meanings, the Croatian and Slovene prefix *od-* and the preposition *od* have the characteristics of Path elements, which have the meaning of direction or the path of trajector's motion with regard to landmark when used with the basic verb *ići* / *iti* 'go', e.g. *Ana* (TR) *je otišla od kuće* (LM) / *Ana* (TR) *je odšla od hiše* (LM) lit. 'Ana (TR) went from house (LM)'; 'Ana (TR) left the house (LM)'.

4 A corpus-based analysis of the Path elements of the verb *otići* / *oditi*

Where there is a 'from'-type Path as there is with the verb *otići* / *oditi*, a correlation is expected between the prefix and the homonymous preposition. However, in addition to repetitive collostructs including the preposition *od* / *od* (*otići od* / *oditi od* 'go from'), the prefixed verb *otići* / *oditi* can also appear with other prepositions as Path elements of motion. On the basis of corpus data, we will attempt to determine the characteristics of prepositional Path elements and establish which element (prefix or preposition) is the dominant Path element marking the direction of motion. In the CQL field, we entered the regular expressions [lemma="otići"][tag="S.*"] into hrWaC and [lemma="oditi"][tag="D.*"] into slWaC, which results in examples of the verb *otići* / *oditi* 'go' followed by any preposition. Typical word order in Croatian and in Slovene is relatively free, which means that the most common word order would be SVO, although other combinations are possible depending on the focus of the sentence. As far as adverbials are concerned, the most common position places the adverb before the verb (*glasno čitati* / *glasno brati* lit. 'loudly read'; 'read aloud'), but when they are in the form of a prepositional phrase, they usually come after the verb (*čitati u sobi* / *brati v sobi* 'read in the room'). This is why our search focused on verb + preposition order and placed the preposition (as the head of a prepositional phrase) in the first place after the verb.

Croatian / Slovene collostructs with *otići* / *oditi* + preposition				
otići u 'go into'	80,548	/	oditi v 'go into'	16,703
otići na 'go onto'	56,708	/	oditi na 'go onto'	14,580
otići sa 'go from'	13,009	/	oditi z 'go from'	5,394
otići do 'go to'	12,193	/	oditi iz 'go out'	3,696
otići iz 'go out'	11,026	/	oditi k 'go to'	2,102
otići kod 'go to'	5,797	/	oditi po 'go for'	1,479
otići po 'go for'	3,313	/	oditi do 'go to'	1,343
otići od 'go from'	2,774	/	oditi od 'from'	1,266
otići za 'go for'	2,017	/	oditi proti 'go to'	1,071
otići k 'go to'	1,758	/	oditi za 'go for'	789

Table 2: 10 most frequent Croatian and Slovene collostructs with *otići* / *oditi* 'leave, depart' + preposition

The search (Table 2) showed that the 10 most frequent Croatian and Slovene collostructs with the verb *otići* / *oditi* + *preposition* correspond highly – 90% of the prepositions are on both lists, and the first three collostructs in Croatian and in Slovene are equivalent: *otići u* / *oditi v* 'go into', *otići na* / *oditi na* 'go onto', *otići sa* / *otići z* 'go from / with'. *Otići od* / *oditi od* is in 8th place in both languages.

Prepositions from the list can be divided into two groups: 'from'-type prepositions (*s* / *z* 'from', *iz* / *iz* 'out of', *od* / *ot* 'from') and 'to'-type prepositions (*u* / *v* 'into', *na* / *na* 'onto', *do* / *do* 'to', *po* / *po* 'for, after', *za* / *za* 'for', *k* / *k* 'at', Croatian *kod* 'at', Slovene *proti* 'towards').

'From'-type prepositions are compatible with the prefixal satellite *od-* / *od-*, which determines *otići* as a 'move from' verb. The preposition *s* / *z* 'from' can appear with both the genitive and instrumental in Croatian and Slovene. With the genitive, it has the meaning of a trajector's departure from a landmark resulting from leaving its surface (*Naravno da nećete otići s otoka ne probajući svježu ribu.* 'Of course you will not leave the island without trying fresh fish.' / *Po skoku boste odšli z letališča bogatejši za izkušnjo...* 'After the jump, you will leave the airport richer for the experience...'). With the instrumental, the parallel in the motion of trajector and landmark is quite similar in nature to the parallel in meaning of the means (Belaj and Tanacković Faletar, 2014:488). A review of examples has shown that some examples have the instrumental after the preposition *s* / *z* 'from' (*Zatim ona ode s volovima na njivu.* 'Then she went into the pasture with the oxen' / *Medtem ko so profesorji odšli z ravnateljico v zbornico, sem ostala z dijaki*. 'While the professors went with the headmaster to the chamber, I stayed with the students.'). Therefore, in the collostructs *otići s* / *oditi z* 'go with', the prepositions *s* / *z* in some of the examples do not bear the features of Path elements of motion – they bear the features of Manner (Talmy, 2007:150) or even Company. The preposition *iz* / *iz* 'out of' has ablative meaning, denoting a trajector's departure from a landmark, but it is initially located in the interior of the landmark, which represents the starting point of its movement and gradual departure (*Trebala bi otići iz ovog stana dok je još na vrijeme.* 'She should leave this apartment while she is still on time.' / *Končno smo prispeli in odšli iz letala na avtobus.* / 'Finally, we arrived and went from the plane to the bus.').

The third of the 'from'-type prepositions in the list is *od* / *ot* 'from'. Prepositional phrases with *od* after a verb of motion are the most obvious elaboration of the concept of source. To elaborate upon the characteristics of source, i.e. of landmark, we searched the corpora for collexemes of the collostruct *otići od* in the range of 0 – 1 and sorted them according to frequency (Table 3).

Croatian and Slovene collexemes with collostructs *otići od* / *oditi od*			
kuće 'house'	566	doma 'home'	674
njega 'him'	275	nas 'us'	76
mene 'me'	176	tam 'there'	45
njih 'them'	120	hiše 'house'	37
nas 'us'	116	tod 'there'	36
nje 'her'	111	njega 'him'	31
tebe 'you'	89	tu 'here'	28
doma 'home'	75	mize 'table'	26
tamo 'there'	74	nje 'her'	25
kuce 'house'	67	mene 'me'	21

Table 3: 10 most frequent Croatian and Slovene collexemes of collostructs *otići od* / *oditi od*

Table 3 shows the 10 most frequent collexemes in the two languages. The resemblances are obvious. Six collexemes in Croatian and four in Slovene are pronouns – the landmark is a person or a group of people. In examples with the singular, the meaning derived is movement away from someone, leaving someone, not only in the physical sense but as a partner, so the metaphoric extension from motion to change of state is present (*...i onda ću otići od njega, od njegovog zanemarivanja* '…and then I'll leave him, his negligence' / *Pogosto se prepirava in takrat si želim, da bi odšla od njega.* 'We often fight and then I want to leave him.'). Examples with *home* (*kuće*, *doma*, and *kuce* – the written form of *kuće* without diacritics – in Croatian, and *doma* in Slovene) can also be interpreted as a metaphoric extension of leaving home and beginning a new phase in life (*…kada sam prvi puta otišao od kuće studirati…* '…when I first left home to study…' / *…zato je nadarjeni mladenič pri dvanajstih letih odšel od doma.* / '…so the gifted young man left home at the age of twelve.'). Examples with adverbs (the Croatian *tamo* 'there', the Slovene *tam* 'there', *tod* 'that way', *tu* 'here') denote moving away from a landmark expressed with a pronominal adverb and addressing a previously mentioned location.

It is possible for *otići* / *oditi* to have two prepositional phrases determining Path: after the prepositional phrase with a 'from'-type preposition, a prepositional phrase with a 'to'-type preposition can appear, especially its antonymous pair (e.g. *od* / *od* 'from'+ *do* / *do* 'to'). The regular expressions [lemma="otići"][word="od"][]{1,3}[word="do"] / [lemma="oditi"][word="od"][]{1,3}[word="do"] extract examples that consist of the verb *otići*, the preposition *od* / *od*, and the preposition *do* / *do* with a distance of one to three words. Of 2,748 examples with *otići od* in hrWac, 46 have this manner of antonymous construction (e.g. *meni je dovoljno da odem od sobe do bicikla i uznojim se...* '... it's enough for me to go from my room to my bike and I get sweaty...'). Some of these examples have temporal meaning or some other meaning, and do not express Path meaning at all. In Slovene, this proportion is 1,262:15 (*S transsibirsko železnico sem odšel od Nakhodke do Moskve.* 'I went from Nakhodka to Moscow by the Trans-Siberian Railway'). We consider such examples regular constructions of 'move from' verbs + 'from'-type prepositions, because the closest position to the verb is considered the focus of information – in this particular example, it is a prepositional phrase with *od* / *od*.

Beside 'from'-type prepositions *s* / *z* 'from', *iz* / *iz* 'out of', *od* / *ot* 'from', all other Croatian and Slovene collostructs with the perfective prefixed verb *otići* / *oditi* contain prepositions with property of Path elements of motion, however, they refer to the goal, which is opposite to the meaning of the prefixal satellite. For example, *u* / *v* 'into' with the accusative[8] expresses the contact directionality of a trajector entering the interior of a landmark (*To sam rekla i otišla u sobu.* 'I said that and went to my room' / *...njihov otrok samozavestno odide v knjižnico...* '...their child goes confidently into the library...'). The preposition *na* / *na* 'onto' comprises the orientation of the movement, the final contact of the trajector and the landmark as a goal, and the inclusion of the landmark as a whole in the process (Belaj and Tanacković Faletar, 2014:425) (*...bio je običaj da svi sudionici otiđu na Zrmanju.* '...it was customary for all participants to go to (the river) Zrmanja.' / *Z velikim navdušenjem smo odšli na London Eye* 'With great enthusiasm, we went to the London Eye.').

Many of these verbs denote motion, but there are also examples with semantic extensions (*Kad je naš sin odlučio otići na liječenje...* 'When our son decided to go for treatment...' / *Leta 2006 je odšel na novo delovno mesto* 'In 2006, he moved to a new job'). This process also enabled some of them to become multiword expressions (*otići u penziju* / *oditi v upokoj* lit. 'to go into retirement', 'to retire'; *otići na put* / *oditi na pot* lit. 'to go on a trip', 'to travel', etc). Some of these can grow new meanings, e.g. the preposition *za* / *za* in the collostructs *otići za* / *oditi za* 'go for' also denotes beginning one's training for a profession (*otići za svećenika* / *oditi za duhovnika* 'become a priest', lit. 'to go for a priest').

The corpus analysis of prefixal-prepositional units in collostructs with the verb *otići* / *oditi* has shown the following: (1) Aside from the repetitive prefixal-prepositional collostruct *otići od* / *oditi od* 'go from', only three of the first 10 most frequent Croatian and Slovene collostructs with the verb *otići* / *oditi* contain a prepositional Path element that denotes distancing or departure (*s* / *z* 'from' with the genitive, *iz* / *iz* 'out of', *od* / *ot* 'from'); (2) Other collostructs contain prepositional Path elements that denote approaching or orientation towards a goal. This meaning is the opposite of the meaning of distancing or departure that appears in the repetitive pattern *otići od* / *oditi od*, which is considered prototypical for this verb (according to Talmy, 1985/2007; Brala-Vukanović and Rubinić, 2011; Brala-Vukanović and Memišević, 2014¹/2014²).

The aforementioned semantic relationships between prefixal and prepositional Path elements with the verb *otići* / *oditi* in Croatian and Slovene indicate that the meaning of the preposition as a Path element of motion neutralizes the meaning of distancing or departing contained in the prefix *od-*

[8] It can also appear with the locative in Croatian and Slovene.

lit. ‘from-’. It also indicates that the property of the Path element of motion weakens in the prefix *od-* in such collostructs, while the property of the perfective aspect prevails.

Unlike prepositions, the prefixes as Path elements in Croatian and Slovene are inseparable from the basic verbs at the level of surface expression. However, at the level of meaning, prepositions appearing in collostructs with some prefixed verbs (like *otići* / *oditi*) represent the key Path element of motion that determines the direction of motion.

5 Conclusion

Using Talmy’s (2007) model, this paper analyzed the Croatian and Slovene verbal prefix *od-* lit. ‘from-’ and the prepositions that follow verbs as Path elements of motion in 10 of the most frequent collostructs appearing in the hrWac and slWaC corpora.

The analysis clarified that Path elements of motion with the verb *otići* / *oditi* in both Croatian and Slovene are expressed via the satellite *od-* / *od-* and a prepositional phrase. Although some theoretical assumptions anticipate that the homonymous prefix-prepositional pattern is the most common, corpus analysis has shown that ‘to’-type prepositions prevail with the ‘move from’ verb *otići* / *oditi*. These results lead to the conclusion that the meaning of perfectivisation in the prefix *od-* / *od-* stands out in relation to the ‘move from’ meaning.

The analysis has also confirmed that two cognate languages, Croatian and Slovene, match at the level of meaning and at the level of surface expression (they have the same morphosyntactic structure). The list of the 10 most frequent prepositions that follow the verb *otići* / *oditi* consists of almost the same prepositions (90%), and the list of collexemes after the collostruct *otići od* / *oditi od* also coincide highly.

The same principle may also be applied to other prefixal-prepositional verbal collostructs in Croatian and Slovene (*do* ‘to’, *na* ‘on’, *u* ‘in’, etc), as well as in other Slavic languages, to determine whether the Path model of motion is universal or if it depends on the semantic potential of the basic verb of motion.

Acknowledgments

The research reported in this paper is supported by the Project MREŽNIK – Hrvatski mrežni rječnik/The Croatian Web Dictionary Project, IP-2016-06-2141 (http://ihjj.hr/projekt/hrvatski-mrezni-rjecnik-mreznik/70/), Hrvatska zaklada za znanost/Croatian Science Foundation (https://epp.hrzz.hr/default.aspx), and Institut za hrvatski jezik i jezikoslovlje/Institute of Croatian Language and Linguistics (http://ihjj.hr/).

References

Stjepan Babić. 1986. *Tvorba riječi u hrvatskom književnom jeziku*: *Nacrt za gramatiku*. Jugoslovenska akademija znanosti i umjetnosti and Globus Zagreb, CRO.

Branimir Belaj and Goran Tanacković Faletar. 2014. *Kognitivna gramatika: Imenska sintagma i sintaksa padežđa, book 1.* Disput, Zagreb, CRO.

Marija Brala Vukanović and Anita Memišević. 2014[1]. English path verbs: A comparative-contrastive English-Croatian analysis. *Jezikoslovlje*, 15(2-3):173-197. http://hrcak.srce.hr/131352

Marija Brala-Vukanović and Anita Memišević. 2014[2]. The Croatian prefix *od-*: A cognitive semantic analysis of Source. *Russian Linguistics.* 38:89-119.

Marija Brala-Vukanović and Nensi Rubinić. 2011. Prostorni prijedlozi i prefiksi u hrvatskom jeziku. *Fluminensia*, 23(2):21-37. http://hrcak.srce.hr/82447

Tomaž Erjavec, Nikola Ljubešić, and Nataša Logar. 2015. The slWaC Corpus of the Slovene Web. *Informatica*, 39(1): 35-42.

hrWaC 2.2. http://nlp.ffzg.hr/resources/corpora/hrwac/

Ronald Langacker. 1987. *Foundations of Cognitive Grammar*: *Theoretical Prerequisites*, *volume 1.* Stanford University Press. Stanford, CA.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In *Text, Speech and Dialogue 2011*, pages 395–402. Springer, Heidelberg http://nlp.ffzg.hr/data/publications/nljubesi/ljubesic11-hrwac.pdf

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9).* Gothenburg: Association for Computational Linguistics, pages 29–35. http://www.aclweb.org/anthology/W14-0405

Ljiljana Mitkovska and Eleni Bužarovska. 2012. The preposition and prefix nad in South Slavic languages with emphasis on Macedonian. *Jezikoslovlje*,13/1):107150. http://hrcak.srce.hr/86247

RHJ = *Rječnik hrvatskoga jezika*. 2000. LZ "Miroslav Krleža" and Školska knjiga, Zagreb, CRO.

slWaC 2.1. http://nlp.ffzg.hr/resources/corpora/slwac/

SSKJ = *Slovar slovenskega knjižnega jezika*. 1997. DZS, Ljubljana, SLO.

Anatol Stefanowitsch and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics volume*, 8(2):209-243. http://www.linguistics.ucsb.edu/faculty/stgries/research/2003_AS-STG_Collostructions_IJCL.pdf

Leonard Talmy. 1985. Lexicalization patterns: semantic structure in lexical forms. *Language Typology and Syntactic Description: Grammatical Categories and the Lexicon*, *volume* 3, 57-149. Cambridge University Press, Cambridge, UK. https://pdfs.semanticscholar.org/96f8/e185a95537aa3fc99c74ffb7755f5fad5884.pdf

Leonard Talmy. 2007. Lexical typologies *Language typology and syntactic description, Grammatical categories and the lexicon, volume 3,* Cambridge University Press, Cambridge, UK.

Morphology of MWU in Quechua

Maximiliano Duran

Université Franche-Comté
Besançon. France
duran-maximiliano@yahoo.fr

Abstract

In this work we study the main characteristics of the morphology of MWU in Quechua and their transformations by multi suffixation. We begin by presenting the grammatical agglutinations of the set of nominal and verbal suffixes which are applicable to MWU. We also present a dictionary of MWU and the construction methods to obtain the NooJ grammars enabling us to generate all their possible inflected forms. We also show how these programs may serve to automatically recognize and annotate MWU in a text.

1 Introduction

Idioms, collocations, verb or nominal compounds, named Entities or domain specific terms may all be considered as MWUs[1] . They are notoriously underrepresented in both Quechua grammar and dictionary handbooks. This work tries to contribute to fill in this lack by proposing a first version of a Quechua-Spanish MWU electronic dictionary, which was written as a result of this work and following the patterns of this study. This paper describes in detail the morpho-syntactic patterns of a class of Quechua MWU. It also presents some typical generating grammars constructed with the help of NOOJ[2] (Silberztein 2003, 2015).

This study is part of our ongoing efforts to create ontology-based lexical resources and linguistic applications developed for Quechua NLP and MT[3]. A similar work has been developed by the the SQUOIA project[4] by A. Rios and A. Göhring (2013).

As it was described in a previous work (Duran 2015) an important class of currently used multi word units, in Quechua are collocations of two-PoS[5] units. There are at least 15 categories 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 table 1.

We can see in this table that any noun or any adjective may be duplicated. The form obtained by the duplication of a noun gives:

a. a new noun, meaning of abundance of what indicates the noun:

N-N>N sacha-sacha> a forest (sacha: tree) (N)

b. a superlative adjective.

N-N>A qari-qari> without fear (qari : man) (A) .

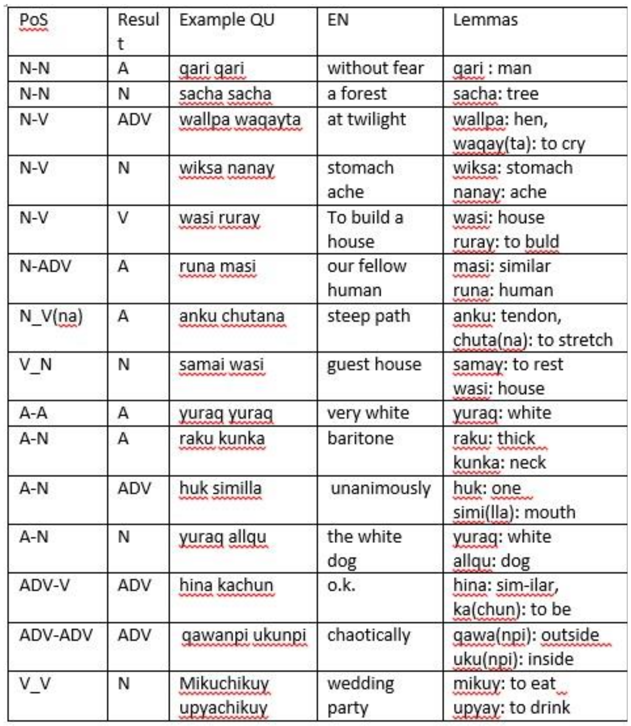
[1] MWUs:Multi word Units.

[2] NooJ is a linguistic development environment software as well as a corpus processor constructed by Max Silberztein.

[3] NLP and MT: Natural Language Processing and Machine Translation.

[4] At the Institute of Computational Linguistics at the University of Zurich, where they have been developed several tools and resources for Cuzco Quechua.

[5] POS part of speech. In this work, we do not include the study of other N-grams.



PoS	Result	Example QU	EN	Lemmas
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-V	N	wiksa nanay	stomach ache	wiksa: stomach nanay: ache
N-V	V	wasi ruray	To build a house	wasi: house ruray: to buld
N-ADV	A	runa masi	our fellow human	masi: similar runa: human
N_V(na)	A	anku chutana	steep path	anku: 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
A-N	ADV	huk similla	unanimously	huk: one simi(lla): mouth
A-N	N	yuraq allqu	the white dog	yuraq: white allqu: dog
ADV-V	ADV	hina kachun	o.k.	hina: sim-ilar, ka(chun): to be
ADV-ADV	ADV	qawanpi ukunpi	chaotically	qawa(npi): outside uku(npi): inside
V_V	N	Mikuchikuy upyachikuy	wedding party	mikuy: to eat upyay: to drink

Table 1. Two-PoS collocatios,in Quechua.

2 The multi-suffixed inflection system

The morpho-syntactic behavior of the different PoS composing a MWU is directly transmitted to the MWU. In fact, we'll see that a MWU inflects in the same manner as each component does. Thus, we shall briefly describe the inflections and derivations of a Noun, an Adjective, an Adverb and a Verb using NooJ[6] grammar.

Nominal inflection. We have in Quechua 68 nominal suffixes Suf-N[7] , devised in two sets:
SUF_N_V = *{-ch, -chá, -cha, -chik, -chiki, -chu, -chu(?), -hina, -kama, -kuna, -lla, -má, -man, -manta, -m, -mpa, -naq, -nta, -ntin, -niraq, -ña, -niq, -p, -pa, -pas, -pi, -poss(7v), -puni, -pura, -qa, -rayku, -raq, -ri, -s, -sapa, -su, -ta, -taq, -wan, -y(!), -ya, -yá, -yupa, -yuq}* (43+7v) regrouping the ones that could be applied to a noun ending in a vowel and
SUF_N_C = {*-chá, -cha, -chik, -chiki, -chu, -chu?, -hina, -kama, -kuna, -lla, -má, -man, -manta, -mi, -nimpa, -naq, -ninta, -nintin, -niraq, -niyuq, -ninka, -ña, -niq, -pa, -paq, -pas, -pi, -poss (7c), -puni, -pura, -qa, -rayku, -raq, -ri, -si, -sapa, -su, -ta, -taq, -wan, -ya(!), -yá, -yupa*} (42+7c) regrouping the ones that can be applied to a noun ending in a consonant.
For instance, for the noun wasi/ house we obtain the genitive wasi-pa/house's, dative wasi-paq/for the house or accusative wasi-ta/the house, cases applying respectively the *-pa*, *-paq* et *-ta* suffixes. All the 50 inflections of a nouns finishing in a vowel, are generated by the following NooJ grammar[8]:

N_V_1 = :CH |:CHAA |:CHIKI |:CHIK |:CHUN |:CHUI |:DCHA |:GEP |:GEPA |:KAMA |:KUNA |:LLA |:MAA |:MAN |:MANTA |:MM |:NAQ |:NTA |:NTIN |:NIRAQ |:NKA |:ñA |:NIQ |:PAQ |:PAS |:PI |:POSV_v |:POSV_c |:PUNI |:PURA |:QA |:HINA |:RAIKU |:RAQ |:RI |:SAPA |:SISV |:SU |:TA |:TAQ |:WAN |:YA |:YY |:YAA |:YUPA |:YUQ;

Where GEPA=*pa*/GEN; is the paradigm of genitive,
TA=*ta*/ACC; is the paradigm of accusative,
PAQ=*paq*/ACC; is the paradigm of dative, etc.
It is possible to agglutinate two, three or more nominal suffixes, for example *kuna-pas*, or *nchik-kuna*-pas, as follows:

Suffixes: -*nchik* /POS + p+1, POS stands for possessive
-pas implies inclusion, also
-kuna is the universal noun-pluralizer
Which give the following inflections:
wasi-kuna/the houses; which follows the paradigm KUNA=kuna/PLU; containing one suffix,
wasi-kuna-pas/including the houses; coming from the paradigm KUNAPAS=*kunapas*/PLUINC; which contain two agglutinated suffixes,

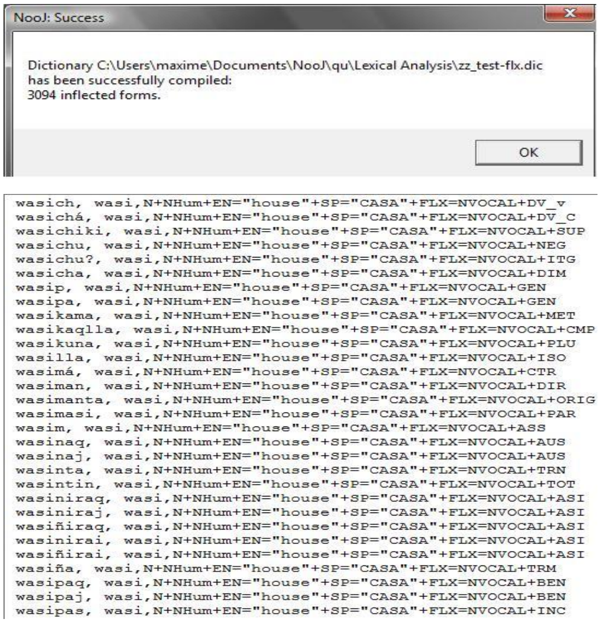
[6] Silberztein, M. [9] NooJ Manual. htpp://www.nooj4nlp.net (220 pages updated regularly).

[7] Suf_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.

[8] For the details of the construction of these paradigms and grammar see Duran (2017) (in printing).

wasi-nchik-kuna-pas/including our houses; coming from the paradigm POSKUNAPAS=*inchikkunapas*/POSPLUINC[9]; containing three suffixes.

For the bi-suffixation of a noun we can obtain a total of 1010 paradigms and for the tri-suffixation 2108 paradigms. In this way we can generate a total of 3094 inflected forms of the noun *wasi*/house containing combinations of less than four suffixes as shown in Fig. 1.



Dictionary C:\Users\maxime\Documents\NooJ\qu\Lexical Analysis\zz_test-flx.dic
has been successfully compiled:
3094 inflected forms.

```
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
wasinaj, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+AUS
wasinta, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+TRN
wasintin, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+TOT
wasiniraq, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+ASI
wasiniraj, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+ASI
wasiñiraq, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+ASI
wasinirai, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+ASI
wasiñirai, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+ASI
wasiña, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+TRM
wasipaq, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+BEN
wasipaj, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+BEN
wasipas, wasi,N+NHum+EN="house"+SP="CASA"+FLX=NVOCAL+INC
```

Fig. 1. A sample of the 3094 inflected forms of *wasi* containing 1, 2 or 3 suffixes.

3 The MWU Quechua-French dictionary.

Between 2013-2017 we prepared an electronic MWU quechua-french dictionnary[10] containing nearly 1500 entries, with 500 forms having a noun as one of the components. As a rule, there is no particularly reliable criterion applied in the ordering of MU's in a dictionary. In general they are listed under the chosen sense of the single-word lemma. We list them alphabetically, based on the first form that comprises the MWU. We can see two typical entries of this dictionary in the following examples:

Mulli pata, N+TOP+MWU+UNAMB+FR="Village Mulli Pata"+SP= "pueblo de Mulli Pata"+FLX=NVOCAL
Chaka-kunka,A+MWU+UNAMB+FR="rauqe"+SP="ronco"+FLX=AVOCAL

Where TOP stands for toponyme, MWU: multi word unit, N for noun, A for adjective, UNAMB for unambiguous form, FLX=NVOCAL: flexional paradigm of nouns ending in a vowel, FLX=AVOCAL: flectional paradigm of adjectives ending in a vowel.

In generation, a query on the first one, *Mulli pata*, gives us 657 inflected forms containing two suffixes like in the following extract:

```
Mulli patachiki,Mulli
pata,N+TOP+MWU+UNAMB+FR="Village
Mulli Pata"+SP="pueblo de Molle
Pata"+FLX=N-V_1+RESG
Mulli patachik,Mulli
pata,N+TOP+MWU+UNAMB+FR="Village
Mulli Pata"+SP="pueblo de Molle
Pata"+FLX=N-V_1+RESG
Mulli patachu,Mulli
pata,N+TOP+MWU+UNAMB+FR="Village
Mulli Pata"+SP="pueblo de Molle
Pata"+FLX=N-V_1+NEG
Mulli patachu?,Mulli
pata,N+TOP+MWU+UNAMB+FR="Village
Mulli Pata"+SP="pueblo de Molle
Pata"+FLX=N-V_1+ITG
```

And the second one, 1184 inflected forms containing two suffixes.

An adjective (A), a pronoun (PRO) or an adverb (ADV) can be inflected by appropriate subsets of Suf_N. For example the set:

SUF_A-V= *{-ch, -chá, -cha,-chik, -chiki, -chu, -chu?, - hina, -kama, -kuna, -lla, -má, -man, -manta, -m, -naq, -nta, -ntin, -niraq, -niq, -ña, -p, -pa, -paq, -pas, -pi, -puni, poss (7v), -pura, -qa, -rayku, -raq, -ri, -s, -sapa, -su, -ta, -taq, -wan, -yá, -yupa}* (40+7v).

Is the subset of suffixes applicable to adjectives ending in a vowel (47). e.g.: *puka*/red will be inflected as:

[9] For the details of the construction of these paradigms and the upcoming calculations of generation see Duran (2013).

[10] For the details of the construction of these paradigms and grammar see Duran (2017). (In printing).

```
 puka-pa/red's   pa is the suffix
of genitive
puka-ta/the red ta is the suffix
of accusative
puka-paq/for the red  paq is the
suffix of benefactor
```

For a pronoun, the respective subset of suffixes contains 36 elements. The following NooJ grammar

PROVOCAL_S =<E>/PRO |:PRO_V-S_1 |:PRO_V_2 |:PRO_V_3;[11]

will generate 515 inflected forms for a singular pronoun finishing in a vowel as is shown below for the pronoun *ñuqa/*I:

```
      ñuqa-pa/mine,
      ñuqa-ta/to me,
      ñuqa-paq/for me.
     ñuqapachá,ñuqa,PRO+FR="je"+
FLX=PRO_V_2+GEN+PRO
      ñuqapahik,ñuqa,PRO+FR="je"+
FLX=PRO_V_2+GEN+RESG1
     ñuqahinachu?,ñuqa,PRO+
FR="je"+FLX=PRO_V_2+CMP+ITG
     ñuqapachu,ñuqa,PRO+FR="je"+
FLX=PRO_V_2+GEN+NEG
      ñuqapahina,ñuqa,PRO+FR="je"+
FLX=PRO_V_2+GEN+CMP
      ñuqapakama,ñuqa,PRO+FR="je"+
FLX=PRO_V_2+GEN+MET
      ñuqapam,ñuqa,PRO+FR="je"+FLX
=PRO_V_2+GEN+ASS
```

For an adverb, the corresponding subset of suffixes contains 34 elements. We may construct the mono suffixed grammar applicable to adverbs ending in a consonant:

ADV_C_1 =:CHÁ |: CHIK |: CHIKI |: CHU |: CHUI |: CHUSINA |: KAMA |: LLA |: MÁ |: MAN |: MANTA |: MI |: NTA |: NINTA |: NINTIN |: NIRAQ |: ÑA |: NIQ |: PAQ |: PAS |: PUNI |: QA |: HINA |: RAQ |: RI |: SI |: TAQ |: WAN |: YÁ |: YUPA;

Some of the inflections of *kunan*/now, generated by this grammar appear below:

```
kunanchá,kunan,ADV+FR="maintenant"
+FLX=ADV-V_1+DINT
kunanchik,kunan,ADV+FR="maintenant
"+FLX=ADV-V_1+RESG1
kunachu,kunan,ADV+FR="maintenant"
+FLX=ADV-V_1+NEG
kunanchu?,kunan,ADV
+FR="maintenant" +FLX=ADV-V_1+ITG
```

4 Multisuffixation of MWU

In this study, we will concentrate our analysis on the inflections of MWUs composed of two categories: N-N, N-V, V-N, A-N, ADV-V.
Let us present some patterns of multi suffix inflections of this kind of MWUs.

1. When we have two nouns N-N, the resulting MWU may be an adjective: *qara uya*/shameless *qara/*leather *uya/*face; or a noun *sacha-sacha*/a forest, where *sacha*/tree. When N-N results an adjective, in general, it expresses a superlative of N: *atuq-atuq*/very smart, where *atuq*/fox.
In these cases, only the second component of the MWU is inflected. The corresponding entries of the dictionary look like as follows:
sacha-sacha, N+MWU+UNAMB+ FR="forêt"+EN=" forest "+FLX=NVOCAL
atuq-atuq, A+MWU+UNAMB+ FR="renard"+EN="very smart"+ FLX=AVOCAL
qara-uya, A+MWU+UNAMB+ FR="effronté"+EN=" shameless "+FLX=AVOCAL
The last two ones will generate 1184 forms like in the following extract:

```
qara uyacha,qara uya,A+MWU
+UNAMB+FR="effronté"+EN="shame-
less"+ FLX=AVOCAL+DIM
```

[11] For its detailed description see Duran(2017a).

```
qara uyap,qara uya,A+MWU
+UNAMB+FR="effronté"+EN="shame-
less"+ FLX=AVOCAL+GEN
qara uyapa,qara uya,A+MWU
+UNAMB+FR="effronté"+EN="shame-
less"+ FLX=AVOCAL+GEN
qara uyantin,qara uya,A+MWU
+UNAMB +FR="ef-
fronté"+EN="shameless"+ FLX=AV-
OCAL+TOT
qara uyahina,qara uya,A+MWU
+UNAMB+FR="effronté"+EN="shame-
less"+ FLX=AVOCAL +CMP
```

2. In the case of MWUs formed by the noun-verb collocation N-V, the resulting expression being a noun N like: *wiksa-nanay*/to have stomach ache, where *wiksa*/stomach, *nanay*/to ache, if the noun inflects, the verb will do so too, but it may happen that only the verb inflects, as in the following examples:
*wiksa-nanaypaq/*it’s for the stomach ache, here, it’s the verb that inflects,
*wiksa-y nana-n/*my stomach aches, here, it’s the noun and the verb that inflect.
anku-chutana/tiresom *anku*/tendron chutay/to stretch, here it’s the verb tat inflects.
This kind of MWU follows the following paradigm:

```
V_IMPE= :SPP1_IMPE |:SIP1_IMPE
|:SIP2_PR_C_IMPE;
```

Which generates 349 inflection like the ones appearing in the following extract:

```
Wiksa-nananchá,wiksa na-
nay,V+MWU+UNAMB+EN="to have
stomach ache"+FR="avoir mal
à l'estomac"+FLX=V_IMPE
+CHPR_IMPE+s+3
wiksa-nananqachu?,wiksa na-
nay,V+MWU+UNAMB+EN="to have
stomach ache"+FR="avoir mal
à l'estomac"
+FLX=V_IMPE+CHUIV_IMPE+s+3
wiksa-nananmá,wiksa na-
nay,V+MWU+UNAMB+EN="to have
stomach ache"+FR="avoir mal
à l'estomac"
+FLX=V_IMPE+CON_IMPE+s+3
wiksa-nanaykachachkan,wiksa
nanay,V+MWU+UNAMB+EN="to
have stomach ache"+FR="avoir
mal à l'estomac"
+FLX=SIP2_PR_C_IMPE
+DISP+PROG+PR+s+3
wiksa-nanaikachachkan,wiksa
nanay,V+MWU+UNAMB+EN="to
have stomach ache"+FR="avoir
mal à l'estomac"
+FLX=SIP2_PR_C_IMPE
+DISP+PROG+PR+s+3
```

3. A MWU of the type V-N may give rise to:
3.1. An imperative expression like: *upallay-simi/silence!* Which does not inflect. or
3.2. An adjective, for instance: *pukllay-siki*/unstable. Where we have the verb *pukllay*/to play and the noun *siki*/ass. Which literally means ‘a playing ass’.
To inflect this MWU, we first nominalize the verb by the agentive suffix *-q* and then we apply the following paradigm:

```
AVOCAL = :A_V_1 |:A_V_2;
```

If we apply the query <AVOCAL> the NooJ engine will generate 1184 MWU inflected forms, containing one or two suffixes, as we show in the following abstract:

```
pukllaq-sikich,pukllaq
siki,A+FR="unstabl"+FLX=AVO-
CAL+DPRO
pukllaq-sikichá,pukllaq
siki,A+FR="unstabl"+FLX=AVO-
CAL+DPRO
pukllaq-sikichik,pukllaq
siki,A+FR="unstabl"+FLX=AVO-
CAL+RESG
pukllaq-sikichu?,pukllaq
siki,A+FR="unstabl"+FLX=AVO-
CAL+ITG
```

```
pukllaq-sikicha,pukllaq
siki,A+FR="unstabl"+FLX=AVO-
CAL+DIM
```

4. The MWU resulting from the collocation of an adjective and a noun A-N gives rise to three possible new categories: an adjective, an adverb or a noun.

4.1.A-N>A *raku-kunka*> baritone (*raku*: thick, *kunka*: neck)

4;2. A-N>ADV *huk-simi(lla)*>unanimously (*huk*: one, *simi*: mouth)

4.3.A-N>N *yuraq-allqu*>the white dog (*yuraq*: white, *allqu*: dog)

Let's see how they inflect.

For 4.1; this MWU will inflect following the adjectival paradigm AVOCAL which will act on the second component even though this one is a noun N. We show below some of these inflected forms:

```
raku-kunkapa,raku
kunka,A+MWU+UNAMB+EN="bari-
tone" +FLX=AVOCAL+GEN
raku-kunkahina,raku
kunka,A+MWU+UNAMB+EN="bari-
tone" +FLX=AVOCAL+CMP
raku-kunkakuna,raku
kunka,A+MWU+UNAMB+EN="bari-
tone" +FLX=AVOCAL+PLU
raku-kunkalla,raku
kunka,A+MWU+UNAMB+EN="bari-
tone" +FLX=AVOCAL+ISO
```

In the case 4.2, the resulting category being an ADV, it will inflect following the adverbial paradigm `ADV_V_1`, defined as follows:

```
ADV_V_1 = :CH |:CHAA |:CHIK
|:CHIKI |:CHU |:CHUI |:CHU-
SINA |:KAMA |: LLA |:MAA |:
MAN |:MANTA |:M |:NTA |:NTIN
|:NIRAQ |:ÑA |:NIQ |:PAQ |:
PAS |:PUNI |:QA |:HINA |:RAQ
|:RI |:SISV |:TAQ |:WAN |:YAA
|:YUPA;
```

This paradigm will influence the second component, even though this one is a noun N. We show an extract of the inflected forms:

```
Huk-simichá,huk
simi,ADV+MWU+UNAMB+EN="unan-
imously" +FLX=ADV-V_1+DINT
Huk-simichu?,huk
simi,ADV+MWU+UNAMB+EN="unan-
imously" +FLX=ADV-V_1+ITG
Huk-similla,huk
simi,ADV+MWU+UNAMB+EN="unan-
imously" +FLX=ADV-V_1+ISO
Huk-simimá,huk
simi,ADV+MWU+UNAMB+EN="unan-
imously" +FLX=ADV-V_1+CTR
Huk-simimanta,huk
simi,ADV+MWU+UNAMB+EN="unan-
imously" +FLX=ADV-V_1+ORIG
```

In the case 4.3, the resulting category being a noun N, it will inflect following the paradigm NVOCAL or NCONSO depending on whtehr the fact that the second component ends in a vowel or a consonant. For instance the MWU: *yuraq-allqu*/the white dog, will generate 657 inflected forms like the following:

```
yuraq-allqucha,yuraq
allqu,N+MWU+UNAMB+FR="le
chien blanc"+EN="the white
dog"+FLX=NVOCAL+DIM
yuraq-allqupa,yuraq
allqu,N+MWU+UNAMB+FR="le
chien blanc"+EN="the white
dog"+FLX=NVOCAL+GEN
yuraq-allqukama,yuraq
allqu,N+MWU+UNAMB+FR="le
chien blanc"+EN="the white
dog"+FLX=NVOCAL+MET
yuraq -allqukuna,yuraq
allqu,N+MWU+UNAMB+FR="le
chien blanc"+EN="the white
dog"+FLX=NVOCAL+PLU
yuraq-allqumanta,yuraq
allqu,N+MWU+UNAMB+FR="le
chien blanc"+EN="the white
dog"+FLX=NVOCAL+ORIG
```

The MWU resulting from the collocation of an adverb and a verb ADV-V>ADV maybe and adverbial expression, like *hina-kachun*/OK (*hina*:/similar, *ka/to be*: to be). It will inflect following the paradigm of its first component ADV, that's to say ADV_V_1 appearing in 4, or ADV_C_1 depending on whether the adverb ends in a vowel or a consonant. For *hina-kachun*, the NooJ parser will generate 17 inflected forms like the following:

```
hinalla-kachun,hina kachun,
ADV+MWU+UNAMB+EN="OK"
+FLX=ADV_V_1_COMP1+ISO
hinaniraq-kachun,hina kachun, ADV+MWU+UNAMB+EN="OK"
+FLX=ADV_V_1_COMP1+AS
hinaña-kachun,hina kachun,
ADV+MWU+UNAMB+EN="OK"
+FLX=ADV_V_1_COMP1+TRM
hinapas-kachun,hina kachun,ADV+MWU+UNAMB+EN="OK"
+FLX=ADV_V_1_COMP1+INC
hinapuni-kachun,hina kachun,
ADV+MWU+UNAMB+EN="OK"
+FLX=ADV_V_1_COMP1+ABS
hinataq-kachun,hina kachun,
ADV+MWU+UNAMB+EN="OK"
+FLX=ADV_V_1_COMP1+CON
hinayá-kachun,hina kachun,
ADV+MWU+UNAMB+EN="OK"
+FLX=ADV_V_1_COMP1+EVD
```

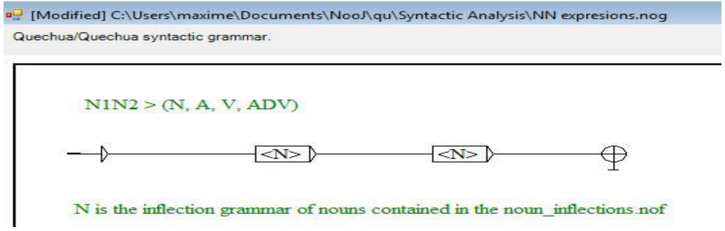
.

5 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 may or may not remain in the same semantic field as one of the components.

yana-uma/traitor (lit. black(A) head(N)) becomes traitor (A).
sunqu-suwa/ a flirt (lit. heart(N) becomes thief(A)) a flirt (A)
piki-piki/ very fast (lit. flea(N) flea) becomes very fast (A)
qallu-qallu/ a parasite of the lever (lit. tongue (N) tongue(N)) becomes a parasite of the lever (N)
kachi-kachi/ dragonfly (lit. salt(N) salt(N)) becomes a dragonfly (N)

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



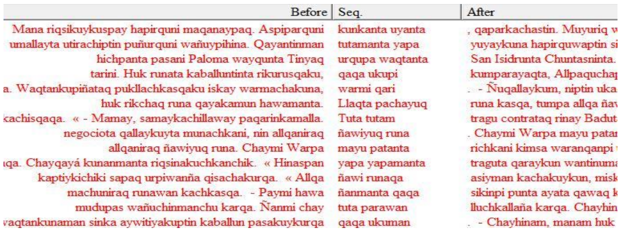
[Modified] C:\Users\maxime\Documents\NooJ\qu\Syntactic Analysis\NN expresions.nog

Quechua/Quechua syntactic grammar.

N1N2 > (N, A, V, ADV)

N is the inflection grammar of nouns contained in the noun_inflections.nof

Fig. 3. This N-N grammar identifies the noun-noun MWUs 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, Allpaquchaq
a. Waqtankupiñataq pukllachkasqaku iskay warmachakuna,	warmi qari	. - Ñuqallaykum, niptin uka
huk rikchaq runa qayakamun hawamanta.	Llaqta pachayuq	runa kasqa, tumpa allqa ñaw
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

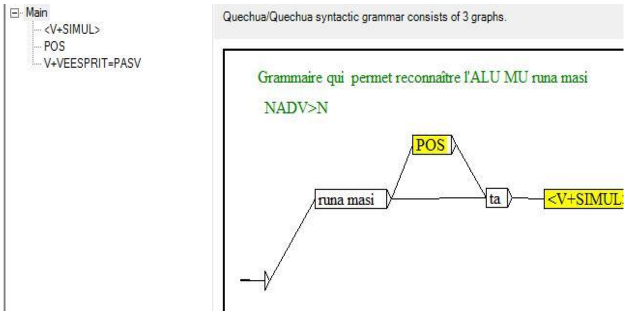
Fig. 4. Output of the Nooj grammar which retrieves N-N multiword units in the corpus

How does it work? In the output sheet, let's take for instance, the third line form: *urqupa waqtanta*/by the mountain's slope; Nooj proposes this collocation because in the inflected dictionary, among the thousands of inflected forms of the noun (N) *urqu/* the mountain, it has found the genitive form *urqupa*/belonging to the mountain, and also the translative form *waqtanta/* by the slope of, corresponding to the adjective (A) *waqta/*side, so it proposes this one.

But, not all of these outputs are actually valid MWUs. For instance, in line 12, Nooj wrongly proposes the form (N1-N2) *ñanmanta qaqa*: (lit. rock from the path) as a MWU. In this case of N1-N2, the rule says that only the second component may be inflected not the first one, so it should not appear here as a MWU.

In our program, we have introduced, this remark as a filter. When we apply a new query the (N1-N2) *ñanmanta qaqa* is not proposed as a MWU anymore. In a similar way the collocation *qaran uyari* is rejected as a MWU candidate by NooJ, because it finds that the first component *qaran*/its leather, its skin, is an inflected form; on the other hand, the same program accepts: *qara-uyanwan/* behaving as a rascal, as a valid inflection.

Thus to avoid over production and ambiguities, it is necessary to introduce disambiguation grammars, based in the Quechua morpho-syntaxis, containing filters like the one in Fig. 5.



Quechua/Quechua syntactic grammar consists of 3 graphs.

Grammaire qui permet reconnaître l'ALU MU runa masi

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

An important resource for handling MWU is to manually make a dictionary of MWUs: a Lexicon-grammar of MWUs, developed following a similar framework as M. Gross [2, 3]. To do this, we have programmed 14 syntactic NooJ grammars like in Fig. 3 one for each category of collocation listed in the introduction, and 12 corresponding disambiguation ones.

We have managed to manually gather more than 1 500 MWU in a lexicon named QU-MWUs. It is the form of an electronic dictionary containing their PoS, the French and Spanish translations, and the accompanying flexional grammar as shown in Fig. 6.

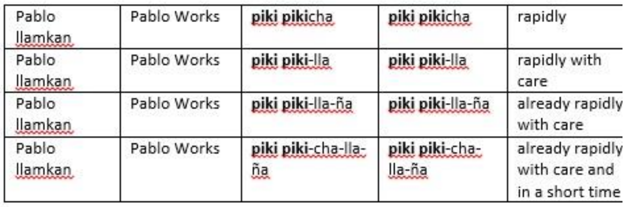
kuyaypaq-kaq, A+MWU+UNAMB +FR="carismatique" +SP=" carismático"+FLX= A_G_1
runa-kay, N+MWU+UNAMB +FR="l'être humain"+SP= "ser humano" +FLX= N_G_1
qepa-punchau, A+MWU+UNAMB +FR=" le passé" +SP="el pasado" +FLX= A_G_1
kuchun-kuchun, A+MWU+UNAMB +FR=" tout les coins « +SP=« de rincon en rincon » +FLX= N_G_1
wallpa-waqayta, N+MWU+UNAMB +FR=" à l'aube «+SP=« al amanecer"+FLX= N_G_1
tawa-chaki, A+MWU+UNAMB +FR=" cuadrupède, bête «+SP=« cuadrúpedo, bestia" +FLX= A_G_1
chulla-ñawi, A+MWU+UNAMB +FR=" bigorne «+SP=« bisco" +FLX= N_G_1
sunqu-suwa, A+MWU+UNAMB +FR=" voleur des coeurs «+SP=« ladron de corazones » +FLX= A_G_1
qaqa-uku, N+MWU+UNAMB +FR=" abïme «+SP=« abismo » +FLX= N_G_1
qinqu-qinqu, A+MWU+UNAMB +FR=" très tordue »+SP=« retorcido » +FLX= A_G_1
ñutuy-ñutuy, A+MWU+UNAMB +FR=" bien molu desmenuzado"+FLX= A_G_1
wicharisqa-intipi, ADV+MWU+UNAMB +FR=" fin matinée «+SP=« tarde en la mañana"+FLX= N_G_1
killan-killan, ADV+MWU+UNAMB +FR="chaque mois"+SP=« mensualmente"+FLX= ADV_G_1
mikuy-maskay,V+MWU+UNAMB +FR="chercher la nouriture"+SP="a la búsqueda de alimentos"+FLX= N_G_1
aqa-wasi, N+MWU+UNAMB +FR="bistrot"+SP=« fonda, bar"+FLX= N_G_1
aka-wasi, N+MWU+UNAMB +FR=" toilette, ambiance où on fait le besoin «+SP=« letrina» +FLX= N_G_1
samai-wasi,N+MWU+UNAMB +FR=" hebergement « +SP=«alojamiento » +FLX= N_G_1
supay-wasi,N+MWU+UNAMB +FR=" l'enfer »+SP=« infierno"+FLX= N_G_1
hurpay-wasi,N+MWU+UNAMB +FR=" hotel "+SP=« hotel, alojamiento"+FLX= N_G_1

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

We expect it to serve us as a linguistic resource in the recognition, the annotation of MWU and in the further project of machine translation technology.

6 Morphology of Quechua MWU sentences

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



Pablo llamkan	Pablo Works	**piki piki**cha	**piki piki**cha	rapidly
Pablo llamkan	Pablo Works	**piki piki**-lla	**piki piki**-lla	rapidly with care
Pablo llamkan	Pablo Works	**piki piki**-lla-ña	**piki piki**-lla-ña	already rapidly with care
Pablo llamkan	Pablo Works	**piki piki**-cha-lla-ña	**piki piki**-cha-lla-ña	already rapidly with care and in a short time

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

The next case concerns a N-ADV MWU: *runa-masi* > human kindness (A) (*runa* (N) *masi* (ADV): similar)
runa-masinchik our fellow human
runa-masinchikta qawaspa kusikuni I am happy seeing our fellow human
runa-masinchikwan kusikuni I am happy with our fellow human
runa-masinchikraiku kusikuni I am happy for our fellow human's sake
Here too, it is the second component: *masi* (ADV) which is inflected. The same grammar in Fig. 3. will generate a large number of transformations of the form *runa-masi*.

We may propose the hypothesis that:

a. 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.
b. if the first component C1 is a verb, the MWU may appear with C1 or C2 inflected
   Example: In the MWU *kuyaypaq-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*;*
   kuyaypaqchus? kanki should I think that you are nice?*; etc.*

I have not yet managed to program the automatic generation of all the valid transformations of this class of MWUs.

7 On the Syntactic Grammar for paraphrase generation involving MWU

The Syntactic Grammar which generates paraphrases/transformations of phrases containing MWUs 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 nor conjunctions, which generally help in the generation of paraphrases. It is the set of suffixes imbedded in the inflections that accomplish these function, 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 [2,3].
It fallows the structure:

N1(m)- N2(pa) C1V,

[12] This example is inspired in that of Simone Vietri [13]

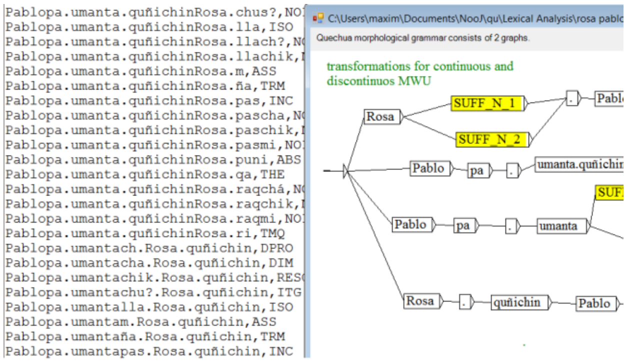
where N1 and N2 represent the free constituents and V, C1 indicate the frozen parts, *-m* and *–pa* are nominal suffixes.

To generate all of the possible paraphrases we have programmed the graphical grammar appearing in Fig. 7. This grammar is formed with 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 grammatically correct 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 a right click 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
Pablopatam umanta quñichin Rosa, …
Rosa-m 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…)
paypa umantam quñichin Rosa
paypatam umanta quñichin Rosa
paymi umanta quñichin payta



```
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 pablo
Quechua morphological grammar consists of 2 graphs.
transformations for continuous and discontinuos MWU
Rosa
SUFF_N_1
SUFF_N_2
.
Pabl
Pablo
pa
.
umanta.quñichin
SUF
Pablo
pa
.
umanta
Rosa
.
quñichin
Pablo

Fig. 7.Sample of paraphrases generated by the graphical grammar

8 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 upcoming projects, we have underlined the need of deploying significant efforts to manually gather more MWUs coming from field work. We have already started gathering a MWU lexicon of more than two components.

Perspectives

We expect to enrich the list of our disambiguation grammars for better MWU recognition.

As mentioned earlier, Quechua remains predominantly an oral language. Thus, an important number of MWUs which are present in the colloquial spoken language do not appear in a written lexicon. We hope to gather more and more of these MWU to integrate them in our QU_MWU dictionary.

We also expect to enhance our method of glossed translations in order to obtain a better automatic translation of MWU into French

References

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

Maximiliano Duran. 2017. *Dictionnaire électronique français-quechua des verbes pour le TAL*. Thèse Doctorale. Université de Franche-Comté. Mars 2017.

Maximiliano Duran. *Formalizing Quechua Noun Inflexion. Formalizing Natural Languages with NooJ*. Edited by A. Donabédian, V. Khurshudian and M.

Silbeztein. Cambridge scholars. Newcastle upon Tyne.

Maximiliano Duran. 2013 *An electronic quechua-french dictionary of MWU*. To be published in Paris in 2017.

Maurice Gross. 1975. Méthodes en syntaxe, Hermann. Paris.

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

Gonçalez Holguin, Diego. 1608. *Vocabulaire de la Lengua General de todo el Perú llamada Lengua Qquichua o del Inca*. Edición y Prólogo de Raúl Porras Barrenechea. Lima, Universidad Nacional Mayor de San Marcos 1952.

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

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

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

Pino Duran, A. German. 1980. *Uchuk Runasimi (Jechua – Quechua).* Conversación y vocabulario Castellano-Quechua Ocopa, Concepción Perú.

Annette Rios et Anne Gôring. 2013. *MachineLearning-Disambiguation of QuechuaVerbMorphology*, in Proceedings of the Second Workshop on Hybrid Approaches to Translation, pages 13–18, Sofia, Bulgaria, August 8, 2013. Association for Computational Linguistics.

Max Silberztein. 2003. *NooJ Manual. 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, pp. 177-190.

Silberztein, M. 2011. *Automatic Transformational* Analysis and Generation, in Proceedings of the 2010 International Nooj Conference, University of Thrace Ed: Komotini, pp. 221-231.

Max Silberztein. 2012. *Variable Unification in* NooJ v3 in Same Volume.

Max Silberztein. 2015. La formalisation des langues. ISTE Editions. London.

Simone Vietri. 2012. *Transformations and Frozen Sentences*, in Proceedings of the 2011 International Nooj Conference: Automatic Processing of various levels of Linguistic Phenomena. Cambridge Scholars Publishing, pp.166-180.