Massively Translingual Compound Analysis and Translation Discovery

Winston Wu, David Yarowsky
Department of Computer Science, Center for Language and Speech Processing
Johns Hopkins University
{wswu,yarowsky}@jhu.edu

Abstract

Word formation via compounding is a very widely observed but quite diverse phenomenon across the world's languages, but the compositional semantics of a compound are often productively correlated between even distant languages. Using only freely available bilingual dictionaries and no annotated training data, we derive novel models for analyzing compound words and effectively generate novel foreign-language translations of English concepts using these models. In addition, we release a massively multilingual dataset of compound words along with their decompositions, covering over 21,000 instances in 329 languages, a previously unprecedented scale which should both productively support machine translation (especially in low resource languages) and also facilitate researchers in their further analysis and modeling of compounds and compound processes across the world's languages.

Keywords: compounds, multilingual, translation

1. Introduction

Morphological compounding (e.g. *lighthouse* or *airport*) is one of the most common and productive methods of word formation across the world's languages (Denning et al., 2007), and yet its derivational processes and semantics can be quite complex.

Consider the semantic concept *hospital*, which can be realized via compound morphology in a remarkable diversity of semantic compositions, including:

Lang.	Compound	Literal Semantics
nl	ziekenhuis	sick + house
no	sykehus	sick + house
hu	kórház	disease + house
eo	malsanuelejo	sick + place
ms	rumah sakit	house + sick
zh	病院	disease + institution

There are clearly a wide variety of semantic associations constituting this concept (e.g. sick/disease + house/place/institution), a variety of constituent orders (e.g. sick+house vs. house+sick) and potentially a variety of compounding processes beyond simple concatenation (e.g. *sykehus* in Norse (no) is a compound of *syk* and *hus* with the insertion of an *e*).

The following paper presents a massively cross-linguistic computational model of both compound morphology compositional processes and compound semantics on a scale of over 300 languages.

Furthermore, the paper not only presents a derived analysis of the compounding process and semantics of compounds within a *single* language (e.g. German), as with much prior related work (e.g. Koehn and Knight (2003)), but does so via a joint model across essentially all the world's languages with adequate dictionary resources, an unprecedentedly large scale for this class of research, and with significant additional synergistic multilingual power.

In addition, the paper successfully applies these models and results to the valuable application of predicting novel translations of compound words, both to English (e.g. *kórház* → *disease* + *house* → *hospital*) and from English (e.g. *hospital* → *disease* + *house, sick* + *place, etc.* → *kórház etc.*), with valuable applications for translation dictionary expansion and out-of-vocabulary handling in machine translation, again on this uniquely large multilingual scale.

Finally, in conjunction with this paper, we release a novel and uniquely large-scale 329-language, 21,000+ example dataset[1] of these compound morphological analyses and their associated compositional and compound translations, a valuable resource for training models for derivational morphology processes and compound semantics, with direct application to machine translation, on this massively multilingual scale.

2. Compound Discovery

We begin only with freely available multilingual translation dictionaries extracted from open-source Wiktionary[2], with the hope that they contain both substantial examples of compounding in each language (e.g. *sykehus* (Norwegian) = *hospital* (English)) as well as translations of the constituents of these compounds (e.g. *syk* = *sick* and *hus* = *house*). Using these dictionaries, we develop a multi-iteration method for discovering both compound translation models or "recipes" motivated across multiple languages that can be used to construct or analyze new compound words that may not be in the dictionary. While there are many existing methods for compound splitting (e.g. Koehn and Knight (2003; Macherey et al. (2011)), we concern ourselves with compounds with two components, which can be combined by simple concatenation, optionally using a glue or filler string at the point of concatenation (Garera and Yarowsky, 2008), or by dropping the last character of the left component (henceforth drop-left). These three compounding processes productively cover a wide spectrum of our multilingual data and serve as an efficient foundation for training the semantic models of compounding in the absence of simple concatenation.

[1] github.com/wswu/worcomal
[2] www.wiktionary.org

Concept		Left +	Right
ninety	zh	九 (nine)	十 (ten)
Monday	nl	maan (moon)	dag (day)
December	fi	joulu (Jule)	kuu (month)
midnight	hu	éj (night)	fél (half)
Frenchman	nl	frans (French)	man (man)
businessman	th	นัก (person)	ธุรกิ (business)
pianist	de	klavier (piano)	spieler (player)
granddaughter	no	datter (daughter)	datter (daughter)
queen	hu	király (king)	nő (woman)

Table 1: Examples of compounds formed by simple concatenation of two components.

In the first iteration of our method, we begin by considering only simple concatenation of two components that both exist in the dictionary (e.g. kór+ház = sick+house). By collecting words in all languages that can be decomposed into such components, we construct compound recipes as in Fig. 1. This process involves accounting for the varied order of components using a reordering and clustering method, augmenting the initial list of compounds in a second iteration of compound discovery by allowing glue characters and the drop-left mechanism, and scoring each word decomposition to indicate its validity as a compound words.
Throughout this paper, we will use the concept "hospital" as a running example. This is an interesting illustrative example since it is not a compound word in English, but occurs as a compound in many other languages.

2.1. Simple Concatenation

Many compound words can be discovered by simply splitting a string into all possible two parts and performing a dictionary lookup on each part. In fact, the large majority of compound words in our dataset are simple concatenations of legal stand-alone words. Table 1 presents a sample of such simple compound words. Note that concatenation can result in false positives (e.g. Dutch hospitaal = hospita 'landlady' + al 'even'), which will be identified by the compound score described later.

2.2. Component Clustering

Within a compound word, the semantic ordering of the components often varies between languages. For example, compound words for the concept "hospital" have different component ordering in different languages:

Dutch:	zieken 'sick'	+	huis 'house'
Malay:	rumah 'house'	+	sakit 'sick'

To account for this variation, we cluster the components[3] using a notion of syntagmatic and paradigmatic analysis. For each concept, we first filter out components that only occur once to remove noise. To illustrate, the top 5 left and right components for the concept "hospital", before correcting for ordering, have the same components on both the left and right sides (Table 2). The numbers indicate the number of languages where we see that component on the left or right side, respectively.

[3]We use components to mean their English translations.

Left		Right	
sick	8	house	7
disease	7	home	6
house	6	institution	4
home	5	place	4
ill	4	court	3

Table 2: Component language counts for "hospital" before correcting for ordering.

hospital =

		+		
sick	11		house	10
disease	7		home	8
ill	7		building	5
illness	5		box	3
sickness	4		family	3
pain	3		place	3
patient	3		case	2
ache	2		dwelling	2
bottle	2		household	2
cottage	2		institution	2

Figure 1: Compounding recipe for the concept 'hospital' using simple concatenation. Numbers indicate the number of languages whose compound words' components had that translation.

Since words in some languages (e.g. Malay) have a "flipped" order relative to the dominant sequence, so "house" and "home" appear both as left and right components. For clustering, we compute a distance matrix between all components, where components on opposite sides of a compound word have a distance of 1, and components on the same side have a distance of 0. Clustering into two clusters with these distances results in cleaner compounding recipes as in Fig. 1. For presentation purposes, we match the order of the recipe components to the most common order across the compound words for a certain concept. Note also that the English component counts are per language, rather than per word. This was done to avoid overcounting (e.g. Hungarian kórház and kórházi both decompose into sick + house, so counting by word would artificially inflate the counts for each of those components).

2.3. Compound Validity Score

Not all words that can be decomposed into valid components are valid compound words. For example, the Dutch hospitaal (decomposed as hospita 'landlady' + al 'even') is clearly not a semantically meaningful compound word. To filter out these poor decompositions, we devise a measure of how well a compound's components follow the compound's recipe. A straightforward but effective score is the geometric mean of the highest counts of the left and right components, respectively. In situations where the a component is not in the recipe, it receives a language count of 0.1 (we do not use 0 because it will zero out the other term in the geometric mean). For example, the Hungarian kórházi (kór + házi; disease + house) receives a score of $\sqrt{8 \cdot 12} = 9.8$, indicating a good decomposition, while the Dutch 'hospitaal' (hospita + al; landlady + even) receives a score of $\sqrt{0.1 \cdot 0.1} = 0.1$, indicating a bad decomposition. Scores for the concept "hospital" are presented in Table 3. Roughly

Score	Lng	Decomposition
12.96	nn	sjukehus = sjuk + hus ; sick + house
12.96	nl	ziekenhuis = zieken + huis ; sick + house
12.96	nl	ziekenhuis = zieke + huis ; sick + house
12.96	tpi	haus sik = haus + sik ; house + sick
12.96	no	sykehus = syk + hus ; sick + house
12.96	nb	sykehus = syk + hus ; sick + house
12.96	ms	rumah sakit = rumah + sakit ; house + sick
12.96	sv	sjukhus = sjuk + hus ; sick + house
12.96	da	sygehus = syge + hus ; sick + house
12.96	da	sygehus = syg + hus ; sick + house
12.96	id	rumah = sakit rumah + sakit ; house + sick
12.96	sv	sjukhus = sjuka + hus ; sick + house
12.96	sv	sjukhus = sjuke + hus ; sick + house
12.96	no	sykehus = syke + hus ; sick + house
10.58	vi	bệnh viện = bệnh + viện ; sick + home
9.79	hu	kórház = kór + ház ; disease + house
9.79	hu	kórházi = kór + házi ; disease + house
6.93	gd	taigh-eiridinn = taigh + eiridinn ; house + patient
6.48	eo	malsanulejo = malsanulo + ejo ; sick + place
4.89	gd	taigh-leighis = taigh + leighis ; house + heal
4.00	zh	病院 = 病癥 + 院; disease + institution
4.00	zh	病院 = 病患 + 院; disease + institution
4.00	zh	病院 = 病 + 院; disease + institution
2.83	zh	病院 = 病者 + 院; patient + institution
2.83	zh	病院 = 病號 + 院; patient + institution
2.83	zh	病院 = 病人 + 院; patient + institution
2.83	zh	病院 = 病體 + 院; sickness + institution
2.00	zh	醫院 = 醫 + 院; heal + institution
1.18	tr	hastane = hasta + ne ; sick + en
1.18	tg	бемористон = бемор + -истон ; sick + -land

Table 3: High-scoring decompositions for compounds of the concept "hospital".

half of the decompositions across all concepts are not valid, and we generally found a score over 2 is sufficient to filter out false positives.

2.4. Compound Augmentation

The first iteration of compound word discovery only took into account simple direct concatenation. However, as noted previously, this is clearly not the only method for forming compounds. To augment the supported processes for generating compound words, we utilize the compounding recipes to construct new words by performing a Cartesian product on the left and right (English) component sets. To construct new compound words, each pair of (foreign) components is concatenated using two new mechanisms: a glue letter, (e.g. Swedish 'construction workers' byggnadsarbetare = byggnad + *s* + arbetare), and dropping the last letter of the left component (e.g. Finnish 'homework' kotitehtävä = koti*o* + tehtävä). We apply component clustering and score the new compound words as previously described. Note that recipes for one concept may result in a compound word for a new concept (e.g. in the second iteration, the Chinese 难处 'difficulty' was constructed using the *hospital* recipe 'ill' (难受) + 'place' (处) using the drop-left mechanism).

For our running example of "hospital", this second iteration resulted in the recipe in Fig. 2. Concatenation resulted in 37 total compounds. The single glue and drop-left increased

hospital =

sick	14		house	12
ill	11		home	8
disease	8		building	5
illness	7		place	4
diseased	5	+	box	3
patient	4		family	3
sickness	4		area	2
pain	3		bottle	2
ache	2		case	2
cure	2		cottage	2

Figure 2: Compounding recipe for the concept 'hospital' including glue character and drop left mechanisms.

this count to 51 compounds, and a 2 character glue added two more words, for a total of 53 compounds. Out of these possible compounds, 17 words had a compounding score less than 1.0, indicating they do not follow the "hospital" recipe.

3. Experiments

We would like to see how well our methods work on compound words it has not seen before. Specifically, we evaluate our compounding methods on two tasks:

1. f2e: Given a foreign compound word, can we decompose and translate it into English?
2. e2f: Given an English concept, can we predict what the compound word would be in a target language?

For these experiments, we randomly chose a test set of 100 languages, with one word from each language that is likely to be a compound word according to our model. Due to the large multilingual breadth of this test set, evaluating on this test set gives a good idea of how the model performs on any given language of the world, rather than focusing on a single language with much more limited cross-linguistic generality. The randomly chosen test set is shown in Table 4. We remove each test word from the training dictionary to simulate it as being out-of-vocabulary.

4. Results and Analysis

For the e2f task, we were able to successfully recover 87 of the 100 test words. In other words, after removing the test word from the dictionary, the model was able to reconstruct the translation of the compound word only from its components, because other languages used the same components in the compounding recipe, either in direct association or via previously unobserved derived associations via the Cartesian product and/or reordering models. This result underscores that even if one does not observe a certain combination of components in any of the training data (e.g. sickness+building), the model's inference that this semantic compounding is viable via model components of both reordering and semantic clustering of observed decompositions of other attested forms translating as *hospital*, facilitates our prediction that this novel association is more likely to occur and mean *hospital* if observed in monolingual target language data.

Lang	Test Pair		Foreign	Literal Foreign	f → e			e → f
Code	English (e)	Foreign (f)	Segmentation	Comp Translations	TopHyp	2ndHyp	Rank	Found?
ace	english	bahsa inggréh	bahsa inggréh	language english	**english**	old english	1	y
ada	thank you	mo tsumi	mo tsumi	you thank	**thank you**	-	1	y
ady	old man	лIыжъы	лIы жъы	man old	**old man**	husband	1	y
af	silkworm	sywurm	sy wurm	silk worm	**silkworm**	-	1	y
akk	head	??	? ?	head go	**head**	begin	1	y
akz	murderer	aatiibi	aati ibi	person kill	homicide	murder	5	y
am	garlic	□□ □□□□□	□□ □□□□□	white onion	**garlic**	-	1	y
ang	dolphin	mereswīn	mere swīn	sea pig	**dolphin**	guinea pig	1	y
arz	there	□□□□	□□□ □	here you	these	here you are	0	y
ast	because	porque	por que	for that	**because**	why	1	y
av	rat	кIудияб гIункIкI	кIудияб гIункIкI	big mouse	**rat**	-	1	n
bg	enviable	завиден	за виден	in eminent	**enviable**	-	1	y
bi	us	yumipela	yu mipela	you us	**us**	virus	1	y
bn	thirteen	□□	□ □	1 three	-	-	0	y
chn	newcomer	cheechako	chee chako	new come	**newcomer**	-	1	y
ckb	newspaper	□□□□□□□□	□□□ □□□□	day letter	agenda	-	0	y
cmn	jade	猪龍	猪龍	hog dragon	**jade**	-	1	y
co	thirteen	trèdeci	trè deci	three ten	**thirteen**	thirty	1	n
crh	swan	aqquş	aq quş	white bird	**swan**	-	1	y
cs	plane angle	rovinný úhel	rovinný úhel	plane angle	**plane angle**	-	1	y
csb	triangle	trzënórt	trzë nórt	three angle	**triangle**	-	1	y
cv	eighteen	вунсаккăр	вун саккăр	ten eight	eighty	eighteenth	0	y
da	cosmodrome	rumhavn	rum havn	space harbour	**cosmodrome**	spaceport	1	y
ee	grandson	tɔgbuiyɔviŋutsu	tɔgbuiyɔvi ŋutsu	grandchild man	**grandson**	-	1	n
enm	housewife	huswif	hus wif	house woman	**housewife**	maid	1	y
eo	yolk	ovoflavo	ovo flavo	egg yellow	**yolk**	egg yolk	1	y
eu	work of art	artelan	arte lan	art work	art	artist	3	y
fi	gene therapy	geenihoito	geeni hoito	gene therapy	**gene therapy**	-	1	y
frm	devil	le diable	le diable	the devil	**devil**	-	1	y
fro	increase	encroistre	en croistre	on increase	**increase**	augment	1	y
fur	cough	tossi	tos si	cough herself	**cough**	-	1	y
ga	consider	déan trácht	déan trácht	do consider	mean	**consider**	2	y
he	pancreas	□□□□	□□ □□	breast heart	heart	**pancreas**	2	y
hi	ten thousand	□□□□□	□□□□ □	thousand zero	**ten thousand**	thousand	1	y
hsb	good evening	dobry wječor	dobry wječor	good evening	**good evening**	good afternoon	1	y
ht	seventy	swasantdis	swasant dis	sixty ten	**seventy**	sixtieth	1	y
hy	anatolia	անատոլիա	անատոլի ա	anatoli oh	**anatolia**	-	1	y
ia	boulder	petra grosse	petra grosse	stone big	**boulder**	capitate bone	1	y
ik	parents-in-law	nulliq-nulliq	nulliq nulliq	mother-in-law father-in-law	**parents-in-law**	-	1	y
io	context	kuntexto	kun texto	with text	**context**	-	1	y
it	saint george	san giorgio	san giorgio	saint george	**saint george**	-	1	y
jbo	eleven	papa	pa pa	one one	-	-	0	y
jv	indonesian	basa indonesia	basa indonesia	language indonesia	**indonesian**	-	1	y
ka	foreign affairs	საგარეო პოლიტიკა	საგარეო პოლიტიკა	foreign politics	**foreign affairs**	foreign policy	1	n
kl	south africa	afrika kujalleq	afrika kujalleq	africa south	**south africa**	-	1	y
km	thirty	□□	□ □	three zero	**thirty**	-	1	y
kpy	eight	ӈыёӷмыллыӈэн	ӈыёӷ мыллыӈэн	three five	**eight**	-	1	y
krl	wristwatch	ranneh̆čuasut	ranneh čuasut	wrist clock	**wristwatch**	-	1	y
li	adverb	biewaord	bie waord	at word	**adverb**	say	1	y
liv	seventeen	seistuoistõn	seis tuoistõn	seven teen	-	-	0	y
lld	twenty-eight	vintot	vint ot	twenty eight	**twenty-eight**	-	1	y
lus	find	hmuchhuak	hmu chhuak	find go	-	-	0	n
mel	leopard	rimau biteang	rimau biteang	tiger star	**leopard**	-	1	y
min	twenty-five	duo puluah limo duo	puluah limo	twenty five	-	-	0	n
mns	fifty	атлов	ат лов	five ten	fifteen	**fifty**	2	y
mt	light year	sena dawl	sena dawl	year light	**light year**	-	1	n
mww	question	lo lus noog lo	lus noog	word ask	-	-	0	y
nb	continent	fastland	fast land	firm country	mainland	**continent**	2	y
nl	abyss	afgrond	af grond	off ground	**abyss**	floor	1	n
nmn	wife	tâa qáe	tâa qáe	person female	female	woman	3	y
nn	arise	oppstå	opp stå	up stand	**arise**	get up	1	y
no	billionaire	milliardær	milliard ær	billion eider	**billionaire**	-	1	y
non	rome	rómaborg	róma borg	rome city	**rome**	-	1	y
nrf	red wine	rouoge vîn	rouoge vîn	red wine	**red wine**	wine	1	y
oc	coal	carbon	car bon	dear good	cheap	**coal**	2	y
ofs	take	nima	ni ma	after one	no	behind	11	y
os	monday	къуырисæр	къуыри сæр	week head	weekend	**monday**	2	y
osx	iron	īsarn	īs arn	ice eagle	**iron**	-	1	y
pih	phonecard	foenkaad	foen kaad	telephone card	**phonecard**	calling card	1	y
pjt	nipple	ipi mulya	ipi mulya	breast face	-	-	0	y
pro	long	lonc tems	lonc tems	long time	duration	long-term	0	y
ps	nine	□□□	□□ □	nine pashto alphabet	**nine**	-	1	n
rm	saddle	sela	se la	up the	**saddle**	-	1	n
ru	dry	высохнуть	вы сохнуть	you dry	wither	section	0	y
sa	garden	□□□□	□□ □□	with forest	submarine	underwater	23	y
sgs	samogitian	žemaitiu kalba	žemaitiu kalba	samogitian language	**samogitian**	-	1	n
sh	give birth	narádati	na rádati	on give birth	bear	bring	11	y
sk	independent	nezávislý	ne závislý	don't addicted	**independent**	-	1	y
sms	june	ǩie´ssmään	ǩie´ss mään	summer month	**june**	-	1	y
sne	whale	kien paos	kien paos	fish whale	**whale**	cetus	1	y
su	silkworm	hileud sutra	hileud sutra	caterpillar silk	**silkworm**	-	1	y
tet	mango	haas-fuan	haas fuan	mango fruit	**mango**	-	1	y
th	lao	□□□□□	□□ □□□	person laos	**lao**	fool	1	y
ti	husband	□□□ □□	□□□ □□	master house	landlord	master	4	y
tpi	bull	bulmakau man	bulmakau man	cow male	**bull**	male	1	y
tzm	not	□□□	□□ □	not and	**not**	grandfather	1	y
udm	russian	ӟуч кыл	ӟуч кыл	russian language	**russian**	old east slavic	1	y
uk	rape	зґвалтувати	з ґвалтувати	from rape	**rape**	-	1	y
uz	dandelion	gulqoqi	gul qoqi	flower dandelion	**dandelion**	-	1	y
vec	author	scritore	scrito re	written king	-	-	0	y
vi	war	chiến tranh	chiến tranh	war fight	**war**	warrior	1	y
vo	seventy	veldeg	vel deg	seven ten	seventeen	**seventy**	2	n
vro	estonian	eesti kiil	eesti kiil	estonia language	**estonian**	-	1	y
wa	somebody	ene sakî	ene sakî	some person	**somebody**	-	1	y
wlm	care	ar ardelw	ar ardelw	on care	defend	listen	0	y
wyi	skull	galk gawang	galk gawang	bone head	**skull**	cranium	1	y
wym	night watchman	nāhtwāhter	nāht wāhter	night guard	**night watchman**	-	1	n
za	zhuang	vahcuengh	vah cuengh	language zhuang	**zhuang**	-	1	y
zh	noise	響聲	響聲	noise sound	sound	**noise**	2	y
zza	step by step	gam gam	gam gam	step stair	**step by step**	-	1	y

Table 4: The 100-language test set (with correct English generation shown in bold).

For the f2e task, we measured both accuracy and mean rank of the model's translation hypotheses. Our method generated the correct English translation in 86/100 cases, a quite respectable performance given the great multilingual diversity of the test set, and the presence of frequently low resource languages where out-of-vocabulary compound words missing from the dictionary are quite common. Out of these cases, the correct English translation had a mean rank of 1.7 in the model's ordered list of hypotheses, indicating that most of the time the correct English trans-

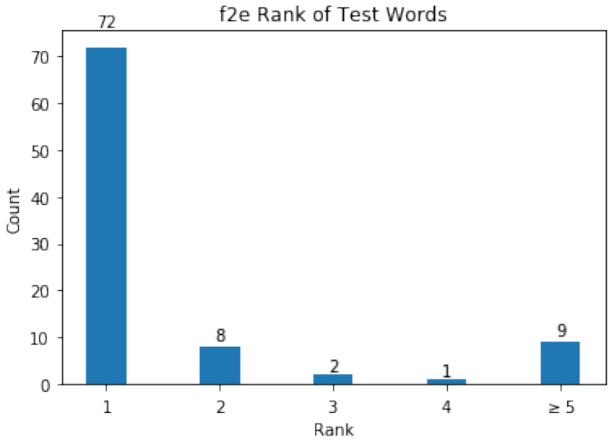

Figure 3: For the f2e task, the top hypothesis for 86/100 test words was correct.

lation was first on the list of hypotheses (Fig. 3), with the "correct" answer (directly matching the test reference) highlighted in bold. We examine a few error cases below.

In cases where the English translation could not be found, the reason was most likely that one or both of the component word translations had not been seen associated with the target English translation in at least one other language, without which it would be impossible to associate and generate the target word. Several test cases were not the top hypothesis (e.g. akz: murderer, he: pancreas, sh: give birth, and several others). However, these errors are quite reasonable (e.g. the predicted translations of akz *aatiibi* = literally *person* + *kill* were *homicide, murder* and *killer (3rd choice)*, which was a reasonable synonym of the reference translation of *murderer*. Likewise, the top hypotheses of the other "errors" are semantically quite similar to their reference translations, even synonyms of the true answer, and indeed could be scored as correct by a manual human evaluation in several cases. A third category of errors (ofs: take, and sa: garden) seem to have occurred because these test words are not actually compound words, and hence not expected to be generated via a compound-morphology-based translation model.

In the case of true unknown words, one would ideally employ our method to generate a large list of possible compound words, then filter them using a monolingual corpus or a language model. However, many of the world's languages are low-resource and do not have broad-coverage monolingual corpora available (excepting a small number of widely-translated works such as the Bible, which have a limited vocabulary). An alternative would be to ask a native speaker to verify the existence of and/or correct choice among the hypothesized these words, a relatively efficient use of native-informant time, especially when prioritized for missing dictionary concepts of high frequency in English and/or high importance in the target domain.

4.1. Dataset Analysis

Utilizing only simple direct concatenation, we were able to discover over 21,000 instances of English concepts that had were direct compounds of simpler constituents of known translation and explainable by one of the model's recipies. By extending the modeled compounding mechanism to a single-character glue or dropping the last character of the left component, our method discovered an additional 2,700 concepts successfully analyzed as compounds.

4.2. Language-Specific Compounding Mechanisms

By examining the different processes used in constructing compound words, we obtain a greater understanding of how specific languages perform compounding. Table 5 shows stereotypical language-specific patterns. For example, most languages construct compounds simply by concatenating two words directly without insertions or deletions (although often in variable order). English often uses 'i' and 's' as glue characters, while German uses 'n' and 's'. This information is not only useful for predicting whether a word is a compound, but can also be useful when generating previously unknown compound word translations into the language. A complete table and statistical analysis of these observed insertions and deletions in each language is included with our dataset, along with language-specific probabilities for the use of each type of compounding mechanism, a useful foundation for any compound-generating language model.

Certain languages like Chinese and Japanese are slightly problematic when discovering compounds using a character-dropping mechanism. For these languages, such a mechanism is not necessarily productive given that the dropped character is not merely a sound-insertion or basic-semantic-linking character but an important semantic component (that would not normally be associated with a single character in an alphabetic or even syllabic writing system). For instance, the Chinese 杀人 murder = 杀害 murder + 人 person is reasonable, but 音乐 music = 音律 tuning + 乐 music is not. Despite this caveat, single character dropping is still a widely observed and productive compounding process in these languages.

5. Related Work

Researchers have explored word compounding, though largely in the monolingual setting or on the order of a couple of languages. One multilingual effort similar to ours is MorBoComp (Guevara et al., 2006), a database of word compounds in 20 languages. The project seems to have stalled, and we were unable to access the data mentioned in their work. Our work encompasses a much larger set of languages (by a factor of 15x) and a much larger set of derived instances (even if their described database was actually available), and posits compound generation and analysis models absent from their work.

While we used straightforward but effective compound splitting algorithms, many more complicated splitting methods have been proposed, e.g. using n-gram counts (Sornlertlamvanich and Tanaka, 1996), supervised methods (Clouet and Daille, 2014), and monolingual and bilingual corpora (Koehn and Knight, 2003; Macherey et al., 2011) and could be productively employed in extensions of our work.

In contrast to several of these other works, the approach and analysis in our paper is simple yet effective in that it only

Lang	Concat	DL	Glue	Common Glues
af	0.75	0.17	0.08	g, s
br	0.85	0.12	0.03	o, r
ca	0.79	0.16	0.05	a, l
co	1.0	0	0	
com	0.5	0.5	0	
cop	0.27	0.73	0	
crh	0.81	0.17	0.01	t, b
cs	0.75	0.21	0.04	o, d
dbl	1.0	0	0	
de	0.8	0.14	0.06	n, s
dv	1.0	0	0	
ee	0.87	0.11	0.02	a
el	0.65	0.35	0	
en	0.93	0.06	0.01	i, s
eo	0.63	0.33	0.05	n, r
es	0.79	0.17	0.04	r, l
esu	1.0	0	0	
et	0.74	0.14	0.12	i, a
eu	0.69	0.24	0.06	k, l
fa	0.55	0.45	0	
fax	1.0	0	0	
ff	1.0	0	0	
fi	0.86	0.1	0.04	n, s
fy	0.5	0.42	0.08	l, t
ha	0.5	0	0.5	n, t
haw	0.57	0.37	0.06	k, h
ht	0.55	0.27	0.18	n, s
hu	0.82	0.14	0.04	i, t
ia	0.83	0.1	0.08	i, l
inh	0.12	0.88	0	
io	0.66	0.24	0.1	a, n
is	0.7	0.22	0.08	a, s
ist	0	1.0	0	
it	0.77	0.18	0.05	s, r
ja	0.32	0.68	0	
jbo	0.21	0.05	0.74	n, r
jv	0.82	0	0.18	n, p
ku	0.7	0.24	0.05	e, d
kum	0	1.0	0	
kw	0.9	0.08	0.02	l
ky	0.93	0.07	0	
la	0.72	0.22	0.06	d, c
lad	0.53	0.29	0.18	g, i
lb	0.76	0.18	0.06	e, s
lv	0.77	0.21	0.02	s, i
pl	0.72	0.22	0.05	o, d
prg	0	1.0	0	
pro	0.55	0.45	0	
ps	0.57	0.43	0	
pt	0.77	0.19	0.04	s, g
qu	0.85	0.14	0.01	y, m
raj	0	1.0	0	
rap	0.88	0.12	0	
rm	0.45	0.43	0.12	r, g
ro	0.77	0.19	0.04	r, i
rup	0.5	0.45	0.05	c, t
scn	0.55	0.33	0.12	n, g
sco	0.48	0.43	0.1	n, g
shn	1.0	0	0	
tpi	0.83	0.14	0.03	k, b
tr	0.85	0.11	0.04	l, s
uz	0.88	0.09	0.03	l, f
vai	1.0	0	0	
vec	0.57	0.32	0.11	n, r
vep	0.71	0.29	0	
vi	0.74	0.25	0.0	t, n
zh	0.34	0.66	0	

Table 5: Percentage of compound words in our dataset that were formed using the each compounding mechanism, along with common glue characters, if applicable.

requires the usually very readily available on-line dictionaries in multiple languages (e.g. via Wiktionary or Panlex) without any analyzed seed training data. Because of this, our approach does not require potentially expensive linguistic annotation, and easily extends to multiple languages, as demonstrated compellingly by our successful scaling to 329 extremely diverse languages incorporating many morphological processes and character sets.

Translation of compound words using dictionaries have been explored by Garera and Yarowsky (2008). Our approach is similar in that we use multiple bilingual dictionaries, but we study and model the compounding phenomenon in more depth as well as on a much, much larger scale, with the significant benefits of much greater novel semantic pair discovery (both via direct observation and via our transitive cluster and reordering models). In addition, we release a very large 329-language 21,000+ instance large public resource of analyzed compound words and components and statistical analyses of their processes across all languages.

In terms of applications, handling compound words well has been shown to improve machine translation, e.g. into English (Koehn and Knight, 2003) and German (Stymne et al., 2013) and has helped simplify medical text (Abrahamsson et al., 2014). We expect that our very large scale publicly distributed compound-based translation dictionaries and associated generative and analytic models will be useful for out-of-vocabulary handling in downstream machine translation systems, especially for low-resource languages.

6. Conclusion

While most languages exhibit broad-scale word formation via compounding, they often differ substantially in terms of the diverse processes by which words compound and novel concepts are realized via these compound processes. Using only freely available bilingual dictionaries and no annotated training data, we derived novel models for analyzing compound words and effectively generated novel foreign-language translations of English concepts using these models. In addition, we release a massively multilingual dataset of compound words along with their decompositions, covering over 21,000 instances in 329 languages, a previously unprecedented scale which we believe will both productively support machine translation (especially in low resource languages) and also facilitate researchers in their further analysis and modeling of compounds and compound processes across the world's languages.

7. Acknowledgments

This work was supported in part by the DARPA LORELEI program. The findings, conclusions, and opinions found in this work are those of the authors and do not necessarily represent the views of the funding agency.

8. Bibliographical References

Abrahamsson, E., Forni, T., Skeppstedt, M., and Kvist, M. (2014). Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language.

Clouet, E. and Daille, B. (2014). Splitting of compound terms in non-prototypical compounding languages.

Denning, K., Kessler, B., and Leben, W. R. (2007). *English vocabulary elements*. Oxford University Press.

Garera, N. and Yarowsky, D. (2008). Translating compounds by learning component gloss translation models via multiple languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Guevara, E., Scalise, S., Bisetto, A., and Melloni, C. (2006). Morbo/comp: a multilingual database of compound words. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA).

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Macherey, K., Dai, A. M., Talbot, D., Popat, A. C., and Och, F. (2011). Language-independent compound splitting with morphological operations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1395–1404. Association for Computational Linguistics.

Sornlertlamvanich, V. and Tanaka, H. (1996). The automatic extraction of open compounds from text corpora. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Stymne, S., Cancedda, N., and Ahrenberg, L. (2013). Generation of compound words in statistical machine translation into compounding languages. *Computational Linguistics*, 39(4).