GENERATING A CONCEPT RELATION NETWORK FOR TURKISH BASED ON CONCEPTNET USING TRANSLATIONAL METHODS

by

Arif Sırrı Özçelik

B.Sc., Statistics, Middle East Technical University, 2010

B.Sc., Computer Engineeering, Middle East Technical University, 2010

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2019

GENERATING A CONCEPT RELATION NETWORK FOR TURKISH BASED ON CONCEPTNET USING TRANSLATIONAL METHODS

APPROVED BY:

Prof. Tunga Güngör ..... (Thesis Supervisor)

Assoc. Prof. Arzucan Özgür .....

Assist. Prof. Şeniz Demir .....

DATE OF APPROVAL:

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor Professor Tunga Güngör at Boğaziçi University for his patience and guidance.

I must express my gratitude to my parents for all the support and continuous encouragement they provided throughout my years of study and the process of writing this thesis.

ABSTRACT

GENERATING A CONCEPT RELATION NETWORK FOR TURKISH BASED ON CONCEPTNET USING TRANSLATIONAL METHODS

ConceptNet is a large-scale network of concepts and relationships, based on various common sense knowledge bases and built upon more than 700 thousand sentences contributed by approximately 15 thousand authors. It was originally developed for the English language and later became a multilingual tool with the addition of other languages using many different sources. It can be seen as a database of how different concepts relate to each other, especially as a valuable resource for systems that perform text analyses, meaning or context extraction. Turkish is a language that lacks similar sources for processing texts and extracting meaning. Although ConceptNet includes examples for Turkish, not many are available where both concepts are in Turkish. This study discusses various methods to create a Turkish ConceptNet using translational techniques based on English ConceptNet and explains the results herewith obtained. Multiple models are tested, using different sources including WordNet, Wikipedia and Google Translate. Results obtained from each model and approaches to improve these results are discussed, while also explaining details, assumptions and drawbacks relevant to each relation.

ÖZET

CONCEPTNET BAZ ALINARAK ÇEVİRİ YÖNTEMLERİYLE TÜRKÇE İÇİN KAVRAMSAL İLİŞKİ AĞI OLUŞTURULMASI

ConceptNet, yaklaşık 15 bin yazarın katkıda bulunduğu 700 binden fazla genel bilgi örneği kullanılarak oluşturulmuş geniş bir kavram ve ilişki ağıdır. Başlangıçta İngilizce için geliştirilmiş olsa da, daha sonra birçok farklı kaynaktan yola çıkılarak diğer dillerin eklenmesiyle çok dilli bir araç haline getirilmiştir. Farklı kavramların birbiriyle olan ilişkilerini barındırması açısından, özellikle metin analizleri, anlam veya bağlam çıkarma işlemleri yapan sistemler için değerli bir kaynak olduğu söylenebilir. Türkçe, metin işleme ve anlam çıkarma sistemleri söz konusu olduğunda diğer dillere kıyasla benzer kaynaklardan yoksun olan bir dildir. ConceptNet, içinde Türkçe veri olmasına rağmen, her iki kavramın da Türkçe olduğu örneklerin sayısı açısından kaynak olarak yeterli değildir. Bu çalışmada, İngilizce ConceptNet baz alınıp çeviri kaynakları kullanılarak, Türkçe ConceptNet oluşturulması amacıyla uygulanmış çeşitli yöntemler tartışılmış ve elde edilen sonuçlar açıklanmıştır. WordNet, Wikipedia ve Google Translate gibi farklı kaynaklar kullanılarak birden fazla model test edilmiştir. Her modelden elde edilen sonuçlar ve bu sonuçları iyileştirmeye yönelik yaklaşımlar tartışılmış, ilişkilerin kendileriyle ilgili detaylar, varsayımlar ve zorluklar açıklanmıştır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZET v
LIST OF FIGURES vii
LIST OF TABLES ix
LIST OF ACRONYMS/ABBREVIATIONS xi
1. INTRODUCTION 1
2. RELATED WORK AND MOTIVATION 5
3. METHODS 11
3.1. Preparations 11
3.2. Tools Used 13
3.3. Model 1 15
3.4. Model 2 18
3.5. Model 3 20
3.6. Model 4 22
3.7. Model 5 24
3.8. Model 6 24
3.9. Model 7 26
4. EXPERIMENTS AND DISCUSSION 30
4.1. Initial Results and Context Enrichment 30
4.2. Utilizing Google Translate as a Monolingual Dictionary 36
4.3. Results For Model 7 42
5. CONCLUSION 55
REFERENCES 57

LIST OF FIGURES

Figure 3.1. Tureng 14

Figure 3.2. Wiktionary 15

Figure 3.3. Wordreference 16

Figure 3.4. Google Translate 17

Figure 3.5. WordNet 17

Figure 3.6. Model 1 18

Figure 3.7. Model 2 19

Figure 3.8. Model 2 - Query WordNet 19

Figure 3.9. Model 2 - Translate Context Groups 20

Figure 3.10. Model 3 21

Figure 3.11. Model 3 - Query WordNet 22

Figure 3.12. Model 4 23

Figure 3.13. Model 5 24

Figure 3.14. Model 6 25

Figure 3.15. Model 7 27

Figure 3.16. Model 7 - Translate & Align Wikipedia Articles 28

Figure 3.17. Model 7 - Score Article Translation Pairs 28

Figure 4.1. Best Performing Relations 48

Figure 4.2. Worst Performing Relations 49

Figure 4.3. Estimated Accuracies vs Start Concept Size 51

Figure 4.4. Estimated Accuracies vs End Concept Size 51

LIST OF TABLES

Table 1.1. List of relations 3

Table 1.2. List of Turkish relations 4

Table 3.1. Stop words list 12

Table 3.2. POS category examples for relations 13

Table 4.1. Sample results for Model 1 32

Table 4.2. Sample results for Model 2 32

Table 4.3. Model 5 results for MadeOf sample 38

Table 4.4. Model 6.1 results for MadeOf sample / matching term count . . . 38

Table 4.5. Model 5 vs Model 6.1 results for MadeOf sample 39

Table 4.6. Model 6.2 results for MadeOf sample / tf-idf scoring 40

Table 4.7. Model 5 vs Model 6.2 results for MadeOf sample 41

Table 4.8. Model 7 results for MadeOf 44

Table 4.9. Model 7 results for SymbolOf 44

Table 4.10. Model 7 results for all relations 46

Table 4.11. Grouped results for all relations 49

Table 4.12. Estimated accuracies vs concept lengths 53

LIST OF ACRONYMS/ABBREVIATIONS

JSON	Javascript Object Notation
ML	Machine Learning
OMCS	Open Mind Common Sense
POS	Part Of Speech
WSD	Word Sense Disambiguation

1. INTRODUCTION

We spend our daily lives in a world full of textual information. The entire internet, but more specifically web pages, online newspapers, magazines, articles, blogs, emails or sites like Wikipedia [1] are just some of the sources that contain this information which are accessed by millions of people every day. There has been huge progress on mining and utilizing data from textual sources of this kind and on statistical approaches extracting useful information by a number of text processing systems. There have also been attempts to create knowledge bases that can be used by such systems aiding to understand textual data better and have a deeper understanding of the semantics. These knowledge bases are crucial for further progress in developing existing text processing capabilities.

One type of source for extracting deeper semantic knowledge are common sense databases. These are textual sources that express common sense knowledge in simple sentence forms. Examples are “birds have wings”, “the sun is very hot” “candy tastes sweet” and so on. Common sense knowledge covers a huge range of what we know and use in our daily lives. When it comes to text processing, humans naturally and extensively use this source of knowledge in understanding and drawing conclusions. So, to actually capture the semantics of any given text an apriori existence of knowledge bases of this kind would help immensely.

Methods like keyword extraction, syntactic parsing and statistical methods have all been used in textual analyses, but the help of large-scale common sense knowledge bases bring can be critical for many applications. Without common sense, a parser could guess that the sentence “I had a terrible evening” has a negative meaning by spotting the keyword “terrible”, but given the sentence “We practiced terribly hard to achieve our goals”, the same parser would not be able to derive a similar conclusion easily.

ConceptNet [2] is a large-scale network of concepts and relations built initially on common sense databases. Common sense knowledge consists of assertions that people use and communicate during everyday life. These assertions include examples like "birds have wings" or "birds are capable of flying". Common sense databases can thus be helpful in building systems that can disambiguate senses, summarize contexts or extract more accurate information from textual data.

OMCS [3] is a public list of common sense knowledge, developed by the MIT Media Lab, gathered through online contributions. It has around 700 thousand sentences contributed by around 15 thousand authors. Certain derivations of assertions by combining different sentences also added new nodes and edges to the network. Currently, it is a multilingual network that brings together knowledge from sources like:

- OMCS in other languages,
- GlobalMind [4] translations,
- "Games with a purpose" like Verbosity (for English), Nadya.jp (for Japanese) etc,
- Wiktionary [5],
- WordNet [6],
- DBPedia [7] that connects Wikipedia articles semantically,
- Open Multilingual WordNet with dictionary type of knowledge,
- UMBEL [8] to connect to another ontology like OpenCyc [9].

ConceptNet5 [10] currently spans 12.5 million edges which represent 8.7 million assertions connecting 3.9 million multilingual nodes [11]. English is the most represented language by 11.5 million edges including at least one English concept. Most of the well represented languages consist of examples contained in existing OMCS knowledge bases. Wiktionary and data collected by GlobalMind are the main sources for translations of concepts from English to other languages.

Table 1.1 shows all 58 relation types (or type of edges) in the network.

Table 1.1. List of relations.

Antonym	AtLocation	Attribute
CapableOf	Causes	CausesDesire
CompoundDerivedFrom	CreatedBy	DefinedAs
DerivedFrom	DesireOf	Desires
Entails	EtymologicallyDerivedFrom	field
genre	HasA	HasContext
HasFirstSubevent	HasLastSubevent	HasPrerequisite
HasProperty	HasSubevent	influenced
influencedBy	InstanceOf	IsA
knownFor	languageFamily	LocatedNear
LocationOfAction	mainInterest	MadeOf
MemberOf	MotivatedByGoal	notableIdea
NotCapableOf	NotCauses	NotDesires
NotHasA	NotHasProperty	NotIsA
NotMadeOf	NotUsedFor	ObstructedBy
participleOf	PartOf	ReceivesAction
RelatedTo	SimilarSize	SimilarTo
spokenIn	SymbolOf	Synonym
TranslationOf	UsedFor	wordnet/adverbPertainsTo
wordnet/adjectivePertainsTo		

The vocabulary of ConceptNet supports a total of 304 languages. To be represented in ConceptNet, a language must have a written orthography, and it should be possible to extract a vocabulary of at least 300 words from ConceptNet's data sources [12].

ConceptNet defines a "Common Language" to be a language included in the network with a vocabulary size of at least 10 thousand terms. There are 68 languages

considered to be a “Common Language” and Turkish is one of them.

Turkish has a vocabulary size of around 66 thousand terms [12]. There are around 10.4 thousand assertions in 7 relations where both concepts are in Turkish. Table 1.2 lists these relations and some examples:

Table 1.2. List of Turkish relations.

Relation	**Size**	**Example**
EtymologicallyDerivedFrom	3073	akdeniz - EtymologicallyDerivedFrom - deniz
RelatedTo	2440	böğürtlen - RelatedTo - ahududu
Synonym	2158	büyülü - Synonym - sihirli
Antonym	1260	fakir - Antonym - zengin
DerivedFrom	1249	kayıkçı - DerivedFrom - kayık
IsA	198	hindi - IsA - kuş
CompoundDerivedFrom	40	emniyet kemeri - CompoundDerivedFrom - kemer

Examples of assertions for English are:

- a bowl - MadeOf - steel
- edinburgh - PartOf - scotland
- brain - UsedFor - think
- sumo wrestler - HasContext - japan
- humanity - Desires - live
- age - Attribute - mature
- karl marx - notableIdea - labor theory of value
- a toy dog - NotIsA - the same as a real dog

2. RELATED WORK AND MOTIVATION

Languages such as English, French, Portuguese have a good amount of sources for ontologies and common sense knowledge, but when it comes to Turkish there is an apparent need for similar sources. Turkish as a language, lacks studies that either create common sense knowledge or somehow translate existing sources in other languages. The second option is what motivated us into implementing the work described in this thesis.

In a study that proposes using common sense data for effective web based distance learning, Anacleto et al. [13] discuss developing and using a OMCS knowledge base for Brazilian Portuguese. A website is setup for users to fill out templates regarding every day life activities. Knowledge collected through this website is later used as part of their study on distance learning.

The ILK (Induction of Linguistic Knowledge) [14] research group at Tilburg University initialized a project in 2008 to develop a OMCS knowledge base for Dutch in collaboration with the developers of OMCS at MIT. This knowledge base was used and further developed by Eckhardt [15] with the help of children playing a web based game.

Balkanet [16, 17] is a collective attempt to create multilingual WordNet lexicons similar to WordNet [18] that spans Greek, Turkish, Romanian, Bulgarian, Czech and Serbian. The result is a large network of synsets that represents semantically related concepts in each individual language and semantically equivalent concepts among these languages. The project makes use of local monolingual WordNets if available, otherwise uses sources like dictionaries, corpora or language specific lexicons. Monolingual Wordnets are developed gradually using independent sources. The process then links each monolingual WordNet to an Inter-Lingual-Index that serves as a centralized index relating synsets among all languages.

Turkish WordNet [19] is a project lead by the Turkish team in the Balkanet project. The initial goal was to create a separate lexicon for Turkish and ultimately inter-connect it with other lexicons at the end. The team started by translating base concepts into Turkish. Later on, a monolingual dictionary was used to extract synonyms, antonyms and hyponyms by rule based methods for these base concepts. In the second phase, they gathered a set of most frequent words in the English language, in what they called a "defining vocabulary". After comparing English words in the defining vocabulary that, when translated, were not in any of the Turkish WordNet synsets, they reached a set of terms that could be searched for synsets in WordNet. These terms were then used to extend the Turkish synset collection through hyperonym-hyponym relations. At the end of this process Turkish WordNet had a size of around 12 thousand synsets with an average synset size of 1.38.

In another study aiming to create a Turkish WordNet named KeNet [20], Yildiz et al. [21] start by extracting synonym candidates from an online dictionary for Turkish. Then they verify synonyms by manually annotating them and create a graph where nodes represent senses connected by synonymy relations. Finally, by looking at clusters they create synsets. They report a larger and more consistent set of results compared to what was obtained by Oflazer et al. [19]. They also mined Turkish Wikipedia for hypernym relations that increased the set of such relations obtained using only a dictionary.

In their study titled SentiTurkNet, Oflazer et al. [22] aimed to create a lexicon of polarity for words in Turkish similar to SentiWordNet for English that could be used by sentiment analysis methods. They used the Turkish version of WordNet [19] by semi-automatically assigning one of the three possible polarity values (positive, neutral and negative) to create SentiTurkNet. In order to do this, they propose a method where all Turkish synsets are assigned a polarity label. Then using a number of features, they train and combine three polarity classifiers in a way that a synset can have scores for each label. The label with the highest score is then accepted as the classification result. What motivated SentiTurkNet was creating a centralized polarity lexicon for Turkish using Turkish WordNet. The methods that created this lexicon are applicable to any

language that has a fairly large synset lexicon. Turkish ConceptNet was inspired by a similar idea to create a concept relation network for Turkish. But unfortunately for Turkish, there is no well established common sense knowledge base to work on.

In their attempt to create a similar common sense list of assertions like what ConceptNet was built on, Ozcan and Amasyali [23] used an online game approach that would ask users to play a game and as a result generate common sense knowledge for Turkish. In this study they look into a number of games previously implemented for English, before implementing a game, as a first version of their system they use translational methods to generate Turkish common sense data, using English ConceptNet, English WordNet, Turkish WordNet and 400 thousand websites as sources. They also utilize a rule based framework [24] to discover concept relations in Turkish using the formal standard Turkish dictionary (Türk Dil Kurumu Sözlüğü) and Turkish Wiktionary. They used Google Translate [25] to translate ConceptNet and WordNet into Turkish but have reported that the result was not reliable due to poor translations. So instead they proposed using a game site called CSOYUN which they kept online for 4 years with 5 different games. These games generate new relations as well as correcting poor translations. They report that 57 thousand reliable concept relations were generated through these online games. Although they also report that around 1.21 million relations included in their database were directly from translations, not much is said how they were translated and how many were corrected. The approach taken in this thesis study uses Google Translate too. However, in contrast to Ozcan and Amasyali [23], we attempted to show that Google Translate could be used as a bilingual online dictionary for English and Turkish, together with Wikipedia articles to disambiguate senses.

Building resources like WordNet for languages that lack apriori lexicons or other knowledge bases can consume much time and resources to accomplish. There have been many studies attempting to create a localized WordNet based on the English WordNet. In one of these studies, Montazery and Faili [26] propose an unsupervised learning approach using a Persian-English bilingual dictionary and a monolingual corpus for training, to generate a Persian WordNet by mapping English synsets to Persian synsets.

An initial small Persian WordNet is built by hand, ML techniques are then used to disambiguate senses in both directions for the rest of the synsets. Links between Persian synsets and English synsets are retained creating an inter lingual WordNet for these two languages. Authors claim the approach can be easily applied to any language given available resources. Another study by the same authors [27] utilizes two bilingual corpora to automatedly generate a Persian WordNet. Word sense disambiguation is then applied based on a score assigned to a synset using these corpora.

Word Sense Disambiguation (WSD) is one of the main problems in computational linguistics. Many application domains including text summarization, concept extraction and machine translation have to solve WSD to some degree in order to produce reliable results. It is also central to the study discussed in this thesis. As each example in ConceptNet relations is short, it can be challenging to be able to select a translation in the target language among many others. In an attempt to tackle the problem of disambiguation where there are low resources and hence ML techniques are not feasible, Sarrafzadeh et al. [28] use Wikipedia as a bilingual corpus. The proposed approach extracts Wikipedia articles in both languages using cross lingual hypertext links that related articles. The authors then use a sense tagger system in the source language and transfer the senses to target language words. This essentially creates a sense tagged bilingual corpus. This corpus can then be used to disambiguate senses while translating between the two languages.

A similar approach was proposed by Sivakumar et al. [29] to solve WSD using Wikipedia link structure and WordNet. Wikipedia articles are related to each other using hyperlinks, which the authors call Interwiki links. If two articles both relate to each other via a hyperlink, the underlying Interwiki link is called a Strong link. Using these links, articles can be logically grouped together under topics. In order to correctly disambiguate a sense for a given word, WordNet synsets are compared against related Wikipedia articles using Lesk's algorithm [30] and Strong links.

Given the assumption that each example in a ConceptNet relation will be small in size and will not carry much information, the problem of translating an example

could be approached in a similar way to how multilingual information retrieval systems consider query translations. In a study investigating the usage of Wikipedia for query translations, Gaillard et al. [31] use categories and cosine similarity scores. Wikipedia articles are organized in a hierarchical manner, similar to WordNet. They initially generate a bilingual dictionary using the Wikipedia French article database and inter lingual links. This initial step generates a list of senses for each article title. Then the query is segmented into chunks (lexical units representing phrases). The best segmentation is determined according to how many chunks there are and how long the chunks are. Less number of chunks which are longer in size are favoured, which the authors call Maximum Forward Matching. Finally, translation candidates in the target language are compared to Wikipedia articles. Using categories, the combination of candidates for all chunks that maximize topic homogeneity is accepted as the correct sense.

Another study by Agirre et al. [32] uses the concept of Conceptual Distance to disambiguate senses, which measures the path between senses in, for example, a hierarchical graph like WordNet. The approach uses subhierarchies in WordNet, considering the other context words in which the term to be disambiguated is placed. A subhierarchy containing a sense of the term together with senses of other context words, positioned deeper in the graph where paths between senses are shorter, overall is regarded having a high Conceptual Density. This measurement takes into account the relative number of senses, making sure densities are not biased. This approach can also be used to disambiguate translations in cases where there are bilingual hierarchical lexicons or for example links between WordNet synsets for two different languages.

Lastly, in their study Speer et al. [33] use ConceptNet together with WordNet to disambiguate senses by utilizing a technique called “blending”. This technique takes a list of senses and coarses some of the glosses or synsets. Common sense data in ConceptNet are merged with synsets from other sources like WordNet through blending which creates a vector space of word senses. This space then can be used to disambiguate words with the help of metrics like cosine similarity.

Following a literature survey of similar areas of research, it becomes apparent that there are not many studies in the domain of common sense databases for Turkish. One of the goals in this study is to offer a contribution for future work especially depending on translations from other languages to Turkish. The other goal is to provide an initial concept relation network that will benefit text processing systems which currently lack semantic knowledge sources of this kind for the Turkish language.

3. METHODS

3.1. Preparations

Before starting translating a relation, the following preparations and assumptions were made:

- Only English to English relations in ConceptNet were considered,
- Nodes on each side of the relation were preprocessed to remove initial stop words like a, an, the, etc.,
- Any translation of a concept to Turkish that fails should result in the assumption that the English concept being translated is a technical or domain specific term so can be accepted as it is into Turkish,
- Except a few specific relations, all concepts were translated in their singularized forms,
- Depending on relations certain Part Of Speech (POS) categories (noun, adjective, verb, etc.) were used to filter senses while translating concepts,
- English terms were lemmatized using the Stanford Core NLP tool [34],
- Turkish terms were lemmatized by Zemberek [35],
- Crawlers were used to extract data from sites like Wiktionary, Wordreference [36], Wikipedia and Google Translate.
- A list of stop words was generated starting with common stop words in English and extending this list after trials when necessary. This list was used to discard certain terms in computing scores like matching term counts. Table 3.1 lists all stop words used.
- All samples used throughout this thesis were randomly selected and evaluated by the author. Accuracy measurements reported were also based on annotations by the author.
- Grammatically incorrect translations accepted as long as meaning is captured.

Table 3.1. Stop words list.

a	can	hasn't	its	ours	there's	what's	%
about	can't	have	itself	ourselves	these	when	especially
above	cannot	haven't	let's	out	they	when's	often
after	could	he	less	over	they'd	where	one
again	couldn't	he'd	lot	own	they'll	where's	two
against	certain	he'll	lot's	part	they're	which	three
all	did	he's	many	previous	they've	while	completely
along	didn't	her	me	possibly	this	who	somewhere
also	do	here	more	same	thing	who's	except
always	does	here's	most	shan't	those	whom	sometimes
am	doesn't	hers	must	she	through	why	within
an	doing	herself	mustn't	she'd	to	why's	around
and	don't	him	much	she'll	too	with	enough
any	down	himself	my	she's	typical	without	ago
another	during	his	myself	should	typically	will	probably
are	easily	how	might	shouldn't	under	won't	actually
aren't	each	how's	may	so	until	would	several
as	e.g.	however	no	some	usually	wouldn't	something
at	eg	I	nor	such	up	whatever	just
be	eg.	I'd	not	small	very	you	i.e.
because	entirely	I'll	next	than	was	you'd	i.e
been	extremely	I'm	of	that	wasn't	you'll	St.
before	few	I've	off	that's	we	you're	St
being	for	if	on	the	we'd	you've	
below	from	in	once	their	we'll	your	
between	further	into	only	theirs	we're	yours	
both	generally	is	or	them	we've	yourself	
but	had	isn't	other	themselves	were	yourselves	
by	hadn't	it	ought	then	weren't	-rrb-	
big	has	it's	our	there	what	's	

Table 3.2 lists some examples of relations and POS categories used for both sides of a relation. Start concept is the concept on the left hand side of a relation and end concept is the concept on the right hand side. Translation candidates that match POS categories are chosen and other categories are discarded.

Table 3.2. POS category examples for relations.

Relation	**Start Concept POS Categories**	**End Concept POS Categories**
MadeOf	noun, adjective, abbreviation, adverb	noun, adjective, abbreviation, adverb
Entails	noun, verb, adjective	noun, adjective, verb
HasContext	noun, adjective, verb, abbreviation	noun, abbreviation
SimilarTo	noun, adjective, abbreviation, verb	noun, verb, adjective

Some of the relations were not included in this study's scope because either they were not suitable, like for example, *TranslationOf* or there were not many examples for English to English relations like *participleOf* or *LocatedNear*.

The text version of ConceptNet, which is in a format similar to JSON, was parsed and all English-to-English assertions were extracted following the assumptions above.

3.2. Tools Used

Dictionaries like Tureng [37], Wiktionary and Wordreference were considered and used as part of different models.

Tureng is an online multilingual dictionary initially built for translations between English and Turkish. It returns many possible translations for a term, including POS categories and certain domains each sense is used in. Figure 3.2 shows results for the term *Bird*.

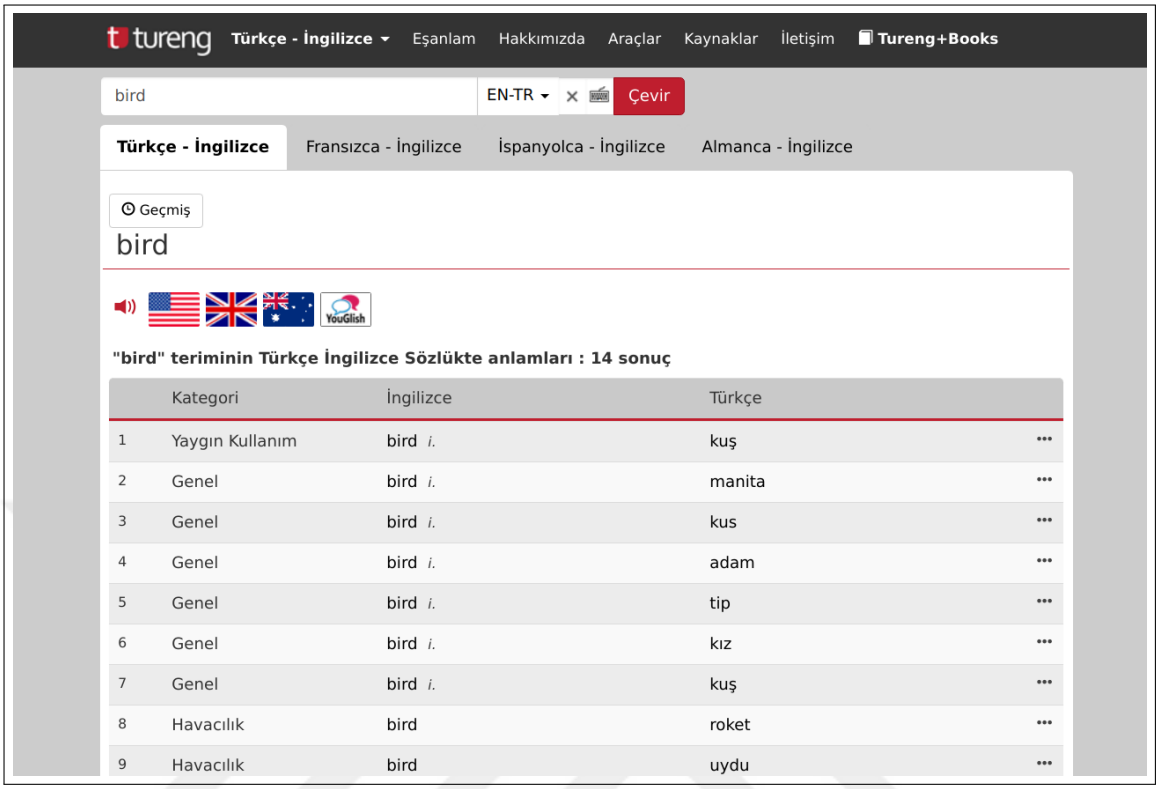


Figure 3.1. Tureng results for *Bird*.

Wiktionary is another online multilingual dictionary owned by the Wikimedia Foundation [38]. Wiktionary entries include glosses, synonyms, examples and statistics for many terms. Turkish Wiktionary accepts queries in English and displays content in both Turkish and English. It contains useful information regarding translations and context for both languages. Figure 3.2 shows results for the term *Bird* on Wiktionary.

Wordreference is an online multilingual dictionary similar to Wiktionary. Figure 3.2 shows results for the term *Bird* on Wordreference.

Google Translate is a publicly accessible online multilingual translation tool. It also provides definitions, examples, synonyms and ranked translations for a term. Figure 3.2 shows results for the term *Bird* returned by Google Translate.



İngilizce [düzenle]

Ad [düzenle]

bird (*çoğulu* **birds**)

Söyleniş [düzenle]

(BK) IPA: **/bɜ:(ɹ)d/**

(20. yüzyılın ortaları New York şehri) IPA: **/bɜjd/**

1. (*omurgalılar*, *hayvan bilimi*, *yiyecekler*) kuş

Örnekler [düzenle]

[1] Ducks and sparrows are **birds**.

[2] He once took in his own mother, and was robbed by a 'pal,' who thought he was a doctor. Oh, he's a rare **bird** is 'Gentleman Joe'!

[3] The door opened and a tall hungry-looking **bird** with a cane and a big nose came in neatly, shut the door behind him against the pre
the desk and placed a wrapped parcel on the desk.

Köken [düzenle]

(bilinmeyen bir köken) → (Eski İngilizce): *bird*, *brid*, *bridd* → (Orta İngilizce)

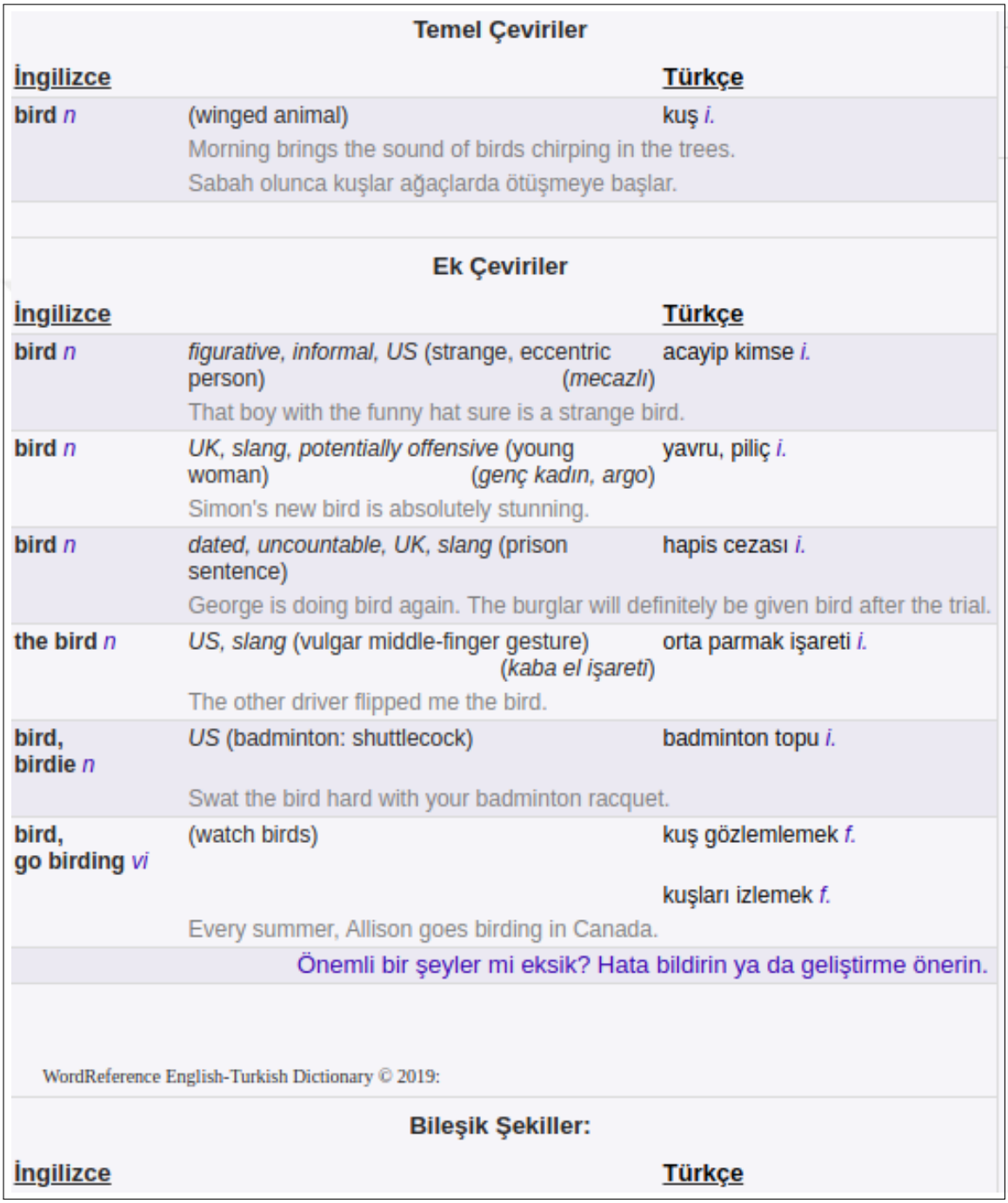
Figure 3.2. Wiktionary results for *Bird*.

WordNet is a lexical database for English that groups synonyms in synsets describing certaing concepts or senses of a term. These synsets are also organized hierarchically into categories. The resulting structure provides a network of senses that are also related through relations like synonymy, hypernymy, hyponynmy, meronymy etc. WordNet is a source for disambiguating terms among different synsets (senses). Figure 3.2 shows results for the term *Bird* on WordNet.

Throughout this study, 7 different models were incrementally developed, each of them attempting to improve certain shortcomings of the previous. Most of these improvements attempts to extend contexts that caused incorrect disambiguation of some terms, resulting in incorrect translations.

3.3. Model 1

Model 1 generates direct translations using Tureng. Figure 3.3 shows pseudo-code of the algorithm. *RELTYPE* represents each relation (type of edges) in the network, like *MadeOf*, *HasA*, *UsedFor* etc. *RELATION* represents examples in each relation. *CONCEPT1* and *CONCEPT2* each represent the concept on the left hand side and right hand side respectively in *RELATION*.



Temel Çeviriler

İngilizce		Türkçe
bird *n*	(winged animal)	kuş *i.*
	Morning brings the sound of birds chirping in the trees.	
	Sabah olunca kuşlar ağaçlarda ötüşmeye başlar.	

Ek Çeviriler

İngilizce		Türkçe
bird *n*	*figurative, informal, US* (strange, eccentric person) (*mecazlı*)	acayip kimse *i.*
	That boy with the funny hat sure is a strange bird.	
bird *n*	*UK, slang, potentially offensive* (young woman) (*genç kadın, argo*)	yavru, piliç *i.*
	Simon's new bird is absolutely stunning.	
bird *n*	*dated, uncountable, UK, slang* (prison sentence)	hapis cezası *i.*
	George is doing bird again. The burglar will definitely be given bird after the trial.	
the bird *n*	*US, slang* (vulgar middle-finger gesture) (*kaba el işareti*)	orta parmak işareti *i.*
	The other driver flipped me the bird.	
bird, birdie *n*	*US* (badminton: shuttlecock)	badminton topu *i.*
	Swat the bird hard with your badminton racquet.	
bird, go birding *vi*	(watch birds)	kuş gözlemlemek *f.* kuşları izlemek *f.*
	Every summer, Allison goes birding in Canada.	

Önemli bir şeyler mi eksik? Hata bildirin ya da geliştirme önerin.

WordReference English-Turkish Dictionary © 2019:

Bileşik Şekiller:

İngilizce	Türkçe

Figure 3.3. Wordreference results for *Bird*.

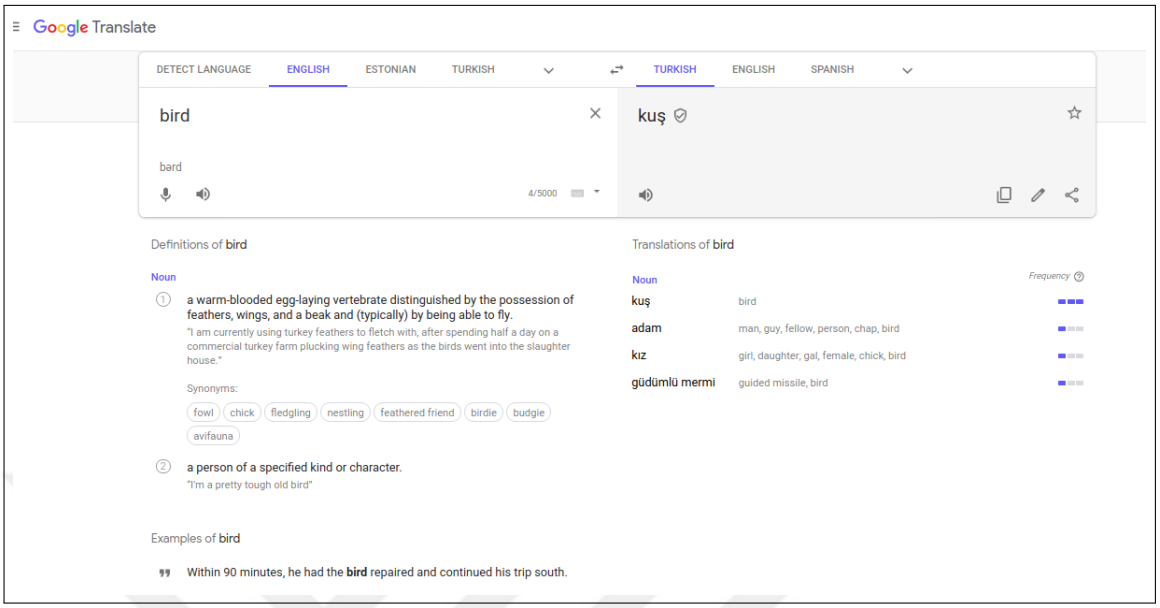


Figure 3.4. Google Translate results for *Bird*.

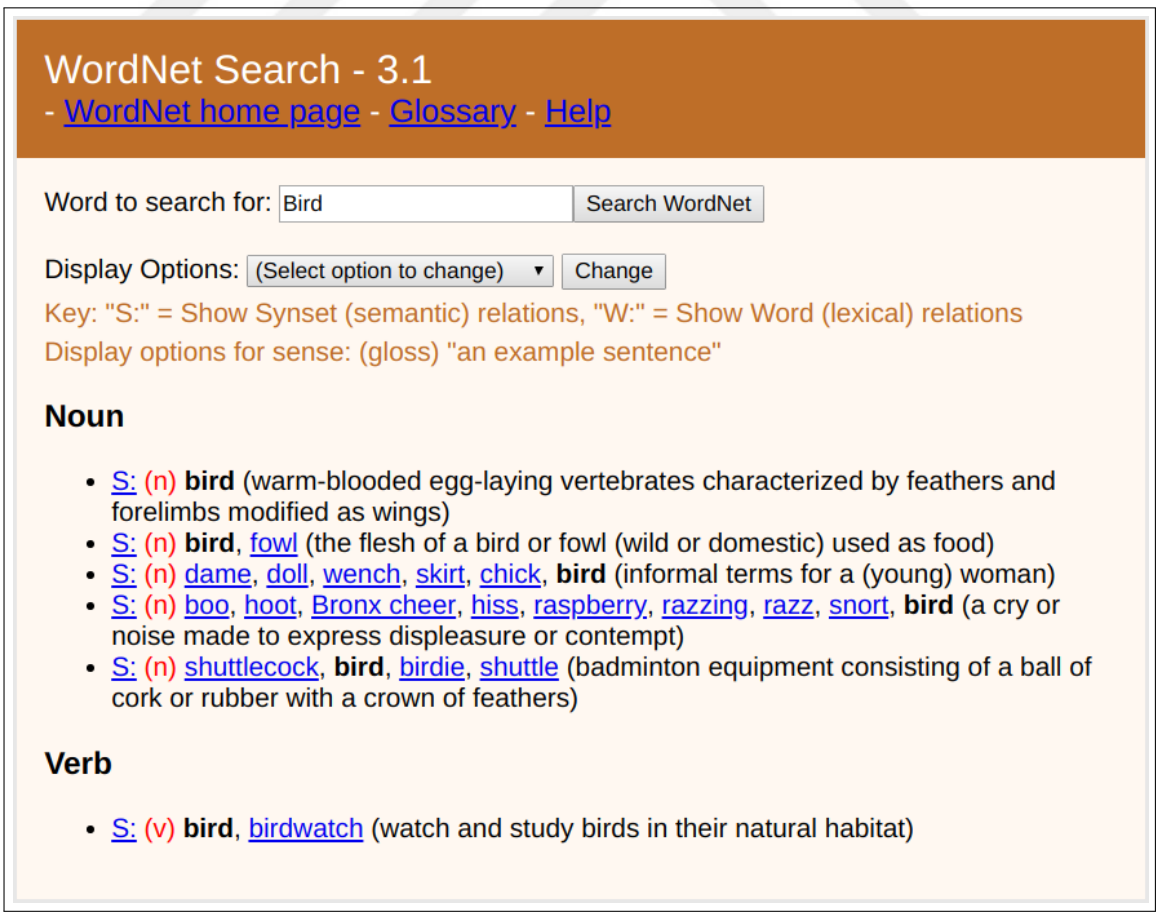


Figure 3.5. WordNet results for *Bird*.

FOR a relation type RELTYPE
FOR an instance of RELTYPE named RELATION
FOR each CONCEPT1 and CONCEPT2 in RELATION:
(i) QUERY Tureng for CONCEPT1
(a) IF translation list is empty RETURN *No Translation*
(b) IF there are translations RETURN the *First Translation* as the accepted translation of CONCEPT1
(ii) REPEAT step (i) for CONCEPT2

Figure 3.6. Model 1.

This model obviously is the most straightforward and simplest approach. Naturally it is expected to result in many incorrect translations as it does not consider any kind of context. Tureng returns many translation candidates for a concept but there is no tool in this approach to say which sense of the translation should be chosen.

3.4. Model 2

Model 2 tries to devise a way to incorporate context before translating concepts. This model parses words and sentences using the Stanford Dependency Parser [34]. Figure 3.4 describes all steps included for this approach. Each entry displayed by Wiktionary is preprocessed and glosses, examples and synonyms are extracted.

WordNet can be used as a context generator before translating concepts. Model 2 also replaces Tureng with Wiktionary. It is not possible to compare and score contexts with Tureng because it generally lacks examples or descriptions for different senses. Wiktionary is a more descriptive online dictionary. However Wiktionary did not prove to be sufficient enough for most of the trial examples either.

FOR a relation type RELTYPE

FOR an instance of RELTYPE named RELATION

FOR each CONCEPT1 and CONCEPT2 in RELATION:

(i) QUERY WordNet as described in Figure 3.4.

(ii) IF no context exists for CONCEPT1, do not translate and flag as *No Translation* and RETURN.

(iii) TRANSLATE *Context Groups* as described in Figure 3.4

(iv) QUERY Turkish Wiktionary for CONCEPT1

  (a) EXTRACT Wiktionary contexts using Wiktionary glosses and examples related to a term.

  (b) COMPARE *lemmatized wiktionary terms* with each *lemmatized context group* generated in step (iii). SCORE the similarity of each *context group* on the basis of shared lemma count.

  (c) ACCEPT the Wiktionary entry with the highest score and accept all the terms of the entry as the accepted translation context

  (d) If all entries score 0, flag as *No Translation* and RETURN.

(v) REPEAT steps (i) through (iv) for CONCEPT2.

Figure 3.7. Model 2.

QUERY WordNet for CONCEPT1 and return all synsets with descriptions.

(i) PARSE each lemma in *synsets* and *descriptions*, look for lemmatized CONCEPT2 in them.

  (a) IF there are hits, return all *Lemmatized Synsets* as *Context Groups* for CONCEPT1.

  (b) If there are no hits, return *No Context* for CONCEPT1.

(ii) IF no synset exists, return *No Context* for CONCEPT1.

Figure 3.8. Model 2 - Query WordNet.

IF there is at least one *Context Group* found in Figure 3.4, translate each lemma using *Model 1* described in Figure 3.3.

(i) IF there are no translations for any of the context groups, flag as *No Translation* and RETURN for CONCEPT1.

(ii) IF there are translations, continue to next step.

Figure 3.9. Model 2 - Translate Context Groups.

3.5. Model 3

Model 3 uses WordReference together with WordNet. Wordreference, similarly to Wiktionary, includes descriptions and terms for both languages, having example sentences, which makes it a great tool for extending contexts. It is a multilingual dictionary, providing links to other dictionaries. It reports keywords in both languages, also displaying examples of usages of versions of the same concept in both languages.

Model 3 also incorporates an augmented version of the Lesk algorithm [30] to enrich WordNet based contexts. The Lesk algorithm disambiguates a word given in a sentence by comparing all synset glosses of this word to synset glosses of other words in the same context. Whichever sense scores the highest number of matching terms, compared to the rest of the glosses for all other words in the window, is chosen to be the correct sense. An example of this algorithm can be seen in the attempt to disambiguate the words in "pine cone", where sense 1 and sense 3 are chosen as the correct senses for pine and cone respectively [30]:

The word pine has two senses:

- Sense 1: kind of evergreen tree with needle–shaped leaves
- Sense 2: waste away through sorrow or illness.

The word cone has three senses:

- Sense 1: solid body which narrows to a point
- Sense 2: something of this shape whether solid or hollow
- Sense 3: fruit of certain evergreen tree

Counting matching terms for the senses above, the Lesk algorithm selects sense 1 for *pine* and sense 3 for *cone*.

Model 3 augments the Lesk algorithm by adding hypernyms, hypernym ancestries, hypoynms and part meronyms to glosses. It translates concepts by disambiguating using WordNet synsets and Wordreference entries. Figure 3.5 describes the approach taken with this model.

FOR a relation type RELTYPE
FOR an instance of RELTYPE named RELATION
FOR each CONCEPT1 and CONCEPT2 in RELATION:
(i) QUERY WordNet as described in Figure 3.5.
(ii) QUERY *Wordreference* for CONCEPT1, retrieve all entries (translation candidate).
(a) For each entry:
i. EXTRACT context using glosses and examples, lemmatize all terms.
ii. COMPARE lemmatized *WordReference* terms to the *Context* generated in step (i). Score the similarity of both contexts on the basis of shared lemma count.
(b) RETURN the entry with the highest score as the correct translation.
(c) IF all entries fail to score, flag as *No Translation* and RETURN.
(iii) REPEAT steps (i) and (ii) for CONCEPT2.

Figure 3.10. Model 3.

QUERY WordNet for CONCEPT1 and return all synsets with lemmatized descriptions, hypernyms, hyponyms and part meronyms up to 6 levels (named *Lemmatized Contexts*).

(i) SEARCH through all lemmas in *Lemmatized Contexts* for lemmatized CONCEPT2.
  (a) IF there are hits, return each *Lemmatized Context* as a *Context Group* for CONCEPT1.
  (b) IF there are no hits, RETURN *No Context* for CONCEPT1.

(ii) IF *No Context* is returned, name CONCEPT1 and CONCEPT2 as the *Context Group* for CONCEPT1.

(iii) JOIN all *Context Groups* to create a single list of lemmas named *Context* for CONCEPT1.

Figure 3.11. Model 3 - Query WordNet.

3.6. Model 4

Model 3 resulted in limited improvement. Concepts on either side of relations in general are small and therefore do not provide enough contextual information to disambiguate using an augmented version of the Lesk algorithm.

Model 4 is actually an algorithm to extend contexts used for translation. The idea is to make use of a parallel corpus [39] consisting of approximately 600 thousand aligned sentences for English and Turkish. Because these sentences are aligned, given a pair of them, some words in the English sentence can be expected to be translations of or at least related to other words in the Turkish sentence. This could be useful in translating words from English to Turkish using only a simple dictionary like Tureng or it could help in extending contexts for concepts (derived from WordNet and Wordreference). Figure 3.6 describes the approach taken in Model 4.

FOR a relation type RELTYPE

FOR an instance of RELTYPE named RELATION

FOR each of the two concepts in RELATION named CONCEPT1 and CONCEPT2

GIVEN a bilingual corpus of English and Turkish sentences (named *Corpus*):

(i) Note all *Context Groups* derived for CONCEPT1 in *Model 3*.

(ii) Note all *Contexts* for entries in *Wordreference* for CONCEPT1 in *Model 3*.

(iii) For each aligned sentence pair in *Corpus*:

  (a) LEMMATIZE English sentences.

  (b) LEMMATIZE Turkish sentences.

(iv) For each lemmatized aligned sentence couple, EN and TR in *Corpus*:

  (a) For each TRANSLATION candidate of CONCEPT1:

    i. COMPARE EN with all terms in each *Context Group* generated in step (i) and *WordReference* terms for TRANSLATION generated in step (ii).

    ii. COMPARE TR with *WordReference* terms for TRANSLATION generated in step (ii).

    iii. IF both comparisons score at least one match, note the sentences and matching terms.

    iv. Using a dictionary like *Tureng* check if matching TR terms are in entries for matching EN terms.

    v. IF there are corresponding matches, EN and TR are candidates to extend *WordReference* contexts (for both English and Turkish) for TRANSLATION.

(v) REPEAT steps (i) through (iv) for CONCEPT2.

Figure 3.12. Model 4.

3.7. Model 5

Model 5 takes a more straightforward and simple approach. After failing to achieve any promising results using previous models, it was decided to use Google Translate as the tool for translating. When a simple query sentence such as "candy is made of sugar" is given as input, Google's translation engine produces a best-possible translation in the target language and also provides a list of candidates for every word in the source query, ranked according to scores (assigned by the same engine). It also provides definitions, examples and synonyms. Figure 3.7 describes the approach taken in Model 5.

FOR a relation type RELTYPE
FOR an instance of RELTYPE named RELATION
FOR each of the two concepts in RELATION named CONCEPT1 and CONCEPT2
USING *Google Translate*:

(i) QUERY *Google Translate* for "CONCEPT1 *isRelatedTo* CONCEPT2".
(ii) COLLECT definitions, examples and first 6 top ranking TRANSLATION candidates.
(iii) RETURN top scoring (assigned by *Google Translate*) candidate.
(iv) REPEAT steps (i) through (iii) for CONCEPT2.

Figure 3.13. Model 5.

3.8. Model 6

Model 6 tries to make use of Wikipedia and cross-lingual links [31] to further correct false positives returned by Model 5. This approach searches for each Turkish translation candidate in Turkish Wikipedia [40], using Google Search API [41], collecting abstracts of the first 10 articles returned. These abstracts are then compared to the Google Translate glosses/definitions and scores are assigned. Translation candidates with the highest scoring abstracts are accepted as the Turkish translations. Also as a

supplementary option, if Google Translate does not return any translation candidates for an English concept, the process falls back to accept the first result reported by Tureng (Model 1). Figure 3.8 describes the approach taken in Model 6.

FOR a relation type RELTYPE

FOR an instance of RELTYPE named RELATION

FOR each of the two concepts in RELATION named CONCEPT1 and CONCEPT2

USING *Google Translate*:

(i) Apply Model 5 steps (i) and (ii).

(ii) IF there are no translations returned, default to *Tureng* and apply *Model 1.* Otherwise continue with next step.

(iii) For each CANDIDATE:

  (a) QUERY *Turkish Wikipedia* using *Google Search API* and retrieve first 10 articles.

  (b) For each ARTICLE:

    i. COMPARE *Google Translate* definition with each ARTICLE abstract. Assign ARTICLE a score based on the number of matching terms.

    ii. SELECT highest scoring ARTICLE abstract as a *Top Ranking Wikipedia Entry.*

  (c) SELECT the CANDIDATE with highest scoring *Top Ranking Wikipedia Entry.*

(iv) REPEAT steps (i) through (iii) for CONCEPT2.

Figure 3.14. Model 6.

3.9. Model 7

Model 6 improves results by using Wikipedia articles and cross lingual links. However not all articles returned by Google Search API have links to a corresponding Turkish article. In any case Wikipedia is a source to be considered in generating contextual information for a relation.

Model 7 uses Google Translate, Wikipedia and Google Custom Search API to select the best possible translation candidate for each concept. It first collects all translation candidates using Google Translate possibly after corrected queries when there are typos. Next it queries the Google Custom Search API with the same query to collect a number of related Wikipedia articles. The first 1500 words are gathered as extracts from each article and then translated into Turkish using Google Translate. Extracts for every article in both languages are the aligned, respecting the same order, in the following manner:

- Every English and Turkish sentence was grouped with the preceding and following sentences.
- These grouped sentences were assumed to be aligned contexts in both languages.
- Preceding and following sentences were used to compensate for one to many mappings between sentences in both languages.

In the final step, for each Google translation candidate of a concept, aligned Wikipedia source and target sentences are scored in the following manner: If at the same time source sentences include the concept and target sentences include the translation candidate, a candidate's score get incremented. This step is repeated for all translation candidates. The translation candidate with the highest scoring Wikipedia article is selected as the translation.

Model 7 makes the assumption that concepts consist of one or two words. In cases where concepts are phrases with more words, Google Translate or Tureng results are accepted directly. Figure 3.9 describes the approach taken in Model 7.

FOR a relation type RELTYPE

FOR an instance of RELTYPE named RELATION

FOR each of the two concepts in RELATION named CONCEPT

(i) IF CONCEPT consists of more than 2 words, QUERY *Google Translate* for "CONCEPT" and RETURN the result as the correct translation. Otherwise proceed to next step.

(ii) APPLY *Model 6* steps (i), (ii) and (iii).

(iii) LEMMATIZE all English terms in CONCEPT. LEMMATIZE all Turkish terms in TRANSLATION.

(iv) QUERY *Google Custom Search API* looking for "CONCEPT1 *is-related-to* CONCEPT2" in *Wikipedia* specifically. COLLECT the first 10 ARTICLES.

(v) TRANSLATE and ALIGN sentences for Wikipedia ARTICLES as described in Figure 3.9.

(vi) ASSIGN scores to ARTICLE - TRANSLATION pairs as described in Figure 3.9.

(vii) CHOOSE the highest scoring ARTICLE - TRANSLATION pair among all and RETURN TRANSLATION as the correct translation.

(viii) REPEAT steps (i) through (vii) for both concepts.

Figure 3.15. Model 7.

For each Wikipedia ARTICLE:

(i) TRANSLATE an extract of the ARTICLE using *Google Translate* including the abstract.

(ii) ALIGN English and translated Turkish sentences for both versions of the ARTICLE.

(iii) For each aligned sentence, EN and TR of ARTICLE:

  (a) LEMMATIZE all English terms in EN.

  (b) LEMMATIZE all Turkish terms in TR.

Figure 3.16. Model 7 - Translate & Align Wikipedia Articles.

For each ARTICLE - TRANSLATION pair:

(i) For each lemmatized aligned sentence, EN and TR of ARTICLE:

  (a) COUNT how many times EN includes CONCEPT and TR includes TRANSLATION.

  (b) Assign the number of times both sentences had matches as the ALIGNED SENTENCE SCORE for TRANSLATION.

(ii) SUM all ALIGNED SENTENCE SCORES to assign the ARTICLE - TRANSLATION pair a score.

Figure 3.17. Model 7 - Score Article Translation Pairs.

Model 7 was selected to be the primary model of translating ConceptNet from English into Turkish, as it proved to be the most promising among all models. The idea that Google Translate could be used for translating Wikipedia article extracts into Turkish with some degree of confidence, helped in creating aligned texts related to each relation. The resulting texts made it possible to extend contextual information in order to disambiguate between different entries.

4. EXPERIMENTS AND DISCUSSION

In this section, results obtained by applying previously mentioned models are revealed. The shortcomings of each model are discussed with examples, suggesting why each model was developed in more detail. Finally accuracy estimations are reported after applying Model 7 to 38 relations.

4.1. Initial Results and Context Enrichment

All models except Model 7 were tested incrementally with randomly selected trial samples from *MadeOf* and *CapableOf*. These trial samples were used to observe results for and improve each model.

Model 1 as expectedly was successful with very straightforward translations while failing to capture context based translations. There were also some results that were not possible to correct due to the treatment of entity names in Tureng, like city or country names.

For example given the concept "Edinburgh" the translation for Turkish should have revealed "Edinburgh" again, while it resulted with "İskoçya'da şehir" which basically just describes Edinburgh as a city in Scotland. The word "bowl" was correctly translated as "kâse" and steel as "çelik". Similarly "organism" as "organizma" and "cell" as "hücre".

An example where context is needed is the relation "brain - UsedFor - think". This should ideally be translated as "beyin - UsedFor - düşünmek or fikir üretmek", but if we only consider nouns and adjectives, the translation of "think" might even end up as "fikir or düşünce". But instead we get "beyin - UsedFor - sanmak". So we need to find a way to incorporate context here.

There are also translations that make no sense like “breathing - UsedFor - meditating” translated as “bir nefeslik zaman - UsedFor - düşünüp taşınmak”. The correct sense for “breathing” should be “soluma or soluk alma or nefes alma”. The sense for “meditating” is actually “meditasyon or meditasyon yapmak” but Tureng does not present that as a possible translation.

Model 2 returned a lot of *No Translation* results for relations that were experimented with, but there were some cases where context did help in returning better translations. For example, the word “wind” can be used both as a verb like in “winding someone up” or as the more commonly used sense like in the sentence “the wind blew the papers off the table”. Model 1 accepts the first case as the accepted translation for “wind” while Model 2 translates wind as “rüzgâr or yel” which is the correct sense in this case. Another example is how the word “jeans” gets translated as “kot” in Model 1. While this is correct, Model 2 adds the term “pantolon” also which generates a further context after translation. A similar example is the word “pencil”, where both models translate it as “kalem”, but Model 2 also discovers “kurşun” as a context.

Most of the *No Translation* results are a result of Wiktionary entries lacking sufficient context terms to use, causing them to be discarded.

At this point the real challenge becomes word sense disambiguation [30, 42, 43]. Making better use of sources like WordNet synsets or Wikipedia article extracts as context groups could be helpful.

An improvement to Model 2 might be to deepen the context generation phase by searching WordNet a second time for the lemmas in relevant synsets concerning a concept. This might reveal more terms attached to context groups. In the case, for example, “glass - MadeOf - silicon” with Model 2, the result is *No Translation* for the concept “glass” but if we look at the synsets of glass the first one will have a description saying “a brittle transparent solid with irregular atomic structure” which will reveal the lemma “atom” which in turn appears in one of the synsets for “silicon”, “silicon, Si, atomic number 14”.

Tables 4.1 and 4.2 below show a comparison of Model 1 and Model 2, reporting both the number of examples with clear incorrect translations and no translations. The comparison includes a small sample of two of the relations in ConceptNet, *MadeOf* and *CapableOf*. These samples were randomly selected examples consisting of both concepts with single terms.

Table 4.1. Sample results for Model 1.

Result	**MadeOf**	**CapableOf**
INCORRECT	25 / 134	38 / 263
NO TRANSLATION (Skip)	6 / 134	9 / 263

Table 4.2. Sample results for Model 2.

Result	**MadeOf**	**CapableOf**
INCORRECT	7 / 134	20 / 263
NO TRANSLATION (Skip)	54 / 134	30 / 263

Model 1 directly uses dictionary entries so as expected has a very low number of examples resulting in *No Translation*. But Model 2 skips a translation if no relevant contextual information can be found so the number of skips increases. Model 2 thus has a lower error rate but higher skip rate.

The following are some examples that Model 1 fails to translate properly:

- Granite – MadeOf – Rock
- Laptop – MadeOf – Chips
- Set – MadeOf – Members
- Computer – CapableOf – Calculate
- Painter – CapableOf – Draw

Model 2 translates “Rock” as “Sallanmak” which is the Turkish version for “to rock a chair” for example. It correctly translates “Granite” but fails to find the correct sense for “Rock”. The first two synsets for the word “Rock” in WordNet are:

- rock, stone (a lump or mass of hard consolidated mineral matter) “he threw a rock at me”
- rock, stone (material consisting of the aggregate of minerals like those making up the Earth’s crust) “that mountain is solid rock”; “stone is abundant in New England and there are many quarries”

So the problem is that Model 2 fails to find the term “Granite” in the above synsets and thus cannot find a context. The synsets for “Granite” on the other hand are:

- granite (plutonic igneous rock having visibly crystalline texture; generally composed of feldspar and mica and quartz)
- granite (something having the quality of granite (unyielding firmness)) “a man of granite”

If instead of comparing the word “Granite” only with synsets for “Rock”, synsets of both words are compared (which is actually the Lesk algorithm), the word “Stone” is added to the context and “Rock” is translated as “Kaya”.

A second example Model 2 fails to translate properly is the word “Chip”. It is translated as “Patates Kızartması” which is “French Fries” in Turkish. There is a similar problem here again as the correct WordNet sense for “Chip” in this example contains words like “Microchip”, “Silicon Chip”, “Microprocessor Chip”, “Electronic, Semiconductor”, “Integrated Circuits”, but there is no mention of “Laptop”, “Computer” or “Personal Computer”. So in order to correctly translate “Chip”, the words “Yonga”, “Kırmık” or “Mikrodevre” are needed.

Synsets for the word “Laptop” in WordNet contain words like “Laptop”, “Computer” and “Portable”. The following is the hypernym relation hierarchy up to 6 ancestors:

- Laptop, Laptop Computer
- Portable Computer
- Personal Computer, PC, Microcomputer (this ancestor node gave us the word Microprocessor)
- Digital Computer
- Computer, Computing Machine etc
- Machine

Traversing the hypernym ancestors introduces words like “Microcomputer”, “Digital”, “Microprocessor”, “Digit”, “Processor” to the context. This results in the correct translation of “Chip”. Another WordNet relation that can be used similarly is the part meronym relation which states the related parts of the concept in consideration. For example “Flat Panel Display” has a part meronym relation with “Portable Computer”. The term “Microprocessor” has also the same relation with “Microcomputer or Personal Computer”. So this relation type is also a good source to look at.

In the third example the word “Set” is actually “Küme” or “Grup” in Turkish. Again similar to “Rock”, the first synset of “Set” refers to the word “Group”, which also is mentioned in the first synset for “Member”.

Model 3 improves Model 2 by augmenting the Lesk algorithm with WordNet hypernyms, part meronym relations and other ancestors up to 6 levels. By using this augmented Lesk algorithm, it was possible to enrich the context of a concept using other related concepts and their senses. While standard Lesk only uses glosses of concepts to score and disambiguate among senses, this model adds different WordNet semantic relations together with glosses and examples of these relations. WordReference was also chosen as the dictionary for translations.

Model 4 resulted in a lot of meaningless matches. Many different domains were used to generate the corpus and these domains are not related to each other. Terms returned by Wordreference were used in unrelated contexts within a domain most of the time. So aligned sentences did not prove to be useful in providing context.

Some of the results achieved for Model 4 using a sample of the *MadeOf* relation with a small extracted parallel corpus [39] on the topic of evolution are (not considering whether the translations are correct):

- Bottle – MadeOf – Glass
  - (i) Concept: Glass
  - (ii) Wordreference Translation: Bardak
  - (iii) En: "Although their eyes are closed for the first week her kittens have no trouble finding the nipples where they can get lifegiving milk—their mother milk—exactly what they need in order to live and grow"
  - (iv) Tr: "İlk hafta gözleri kapalı olmasına rağmen yavrular süt içecekleri yeri bulmakta hiç zorluk çekmezler. Dokuz gün sonra yavruların gözleri açılır Annenin sütü yavruların büyümesi için tam gereken özelliklerdedir"
  - (v) Matching English WordReference context word: Milk
  - (vi) Matching Turkish WordReference context word: Süt
- Lettuce – MadeOf – Water
  - (i) Concept: Water
  - (ii) Wordreference Translation: Su
  - (iii) En: "At exactly this time the mother penguins return from the sea"
  - (iv) Tr: "Tam bu dönemde anne penguenler açık denizden kıyıya dönerler"
  - (v) Matching English WordReference context word: Sea
  - (vi) Matching Turkish WordReference context word: Deniz
- Mountain – MadeOf – Land
  - (i) Concept: Land
  - (ii) Wordreference Translation: Toprak

(iii) En: “It is rich in nutrients and contains some special chemical ingredients that protect the kitten from getting sick”

(iv) Tr: “Her türlü besin açısından zengindir ayrıca yavruyu hastalıklardan koruyan özel bazı kimyasallar da bu sütte bulunur”

(v) Matching English WordReference context word: Rich

(vi) Matching Turkish WordReference context word: Zengin

- People – MadeOf – Meat

(i) Concept: Meat

(ii) Wordreference Translation: Et

(iii) En: “Because of this cooperative system of babysitting other mother giraffes can leave their babies and go kilometers away in search of food”

(iv) Tr: “Bu güvenlik sistemi sayesinde diğer anneler rahatlıkla bebek zürafaları bırakıp kilometrelerce uzağa yiyecek aramaya gidebilirler”

(v) Matching English WordReference context word: Food

(vi) Matching Turkish WordReference context word: Yemek

Words in sentences with matching terms do not necessarily semantically relate to the concept or its translation (like the above resulting match of words “Rich” - “Zengin” for the concept “Land” - “Toprak”). It turned out that the parallel corpus was not very productive in extending contexts for terms, therefore decision was taken to continue with the next model.

4.2. Utilizing Google Translate as a Monolingual Dictionary

Model 5 introduced Google Translate and produced promising results. Using the best scored translation candidate for each concept in the query, this approach surprisingly resulted in around 62% accurate translations excluding technical or domain specific terms that could actually be accepted for both languages. Some successfully translated examples are:

- Granite – MadeOf – Rock (Granite/Granit and Rock/Kaya)
- Laptop – MadeOf – Chips (Chip/Yonga, Laptop could not be translated)
- Set – MadeOf – Members (Member/Üye, Set is translated as Dizi which is questionable)
- Wind – MadeOf – Air (Wind/Rüzgâr and Air/Hava)

There were many *No Translation* and incorrect results. Some examples being:

- Word – MadeOf – Letter (Word / Kelime but Letter / Mektup)
- Scissor – MadeOf – Metal (Scissor / Makas but Metal /Maden)
- Pancake – MadeOf – Milk (Milk / Süt but Pancake / Yassı)
- Computer – MadeOf – Microchip (Computer / Bilgisayar but Microchip could not be translated)
- Table – MadeOf – Tree (Tree / Ağaç but Table / Tablo)

There were also some questionable translations too. For example:

- Mountain – MadeOf – Land (Land /Arazi)
- Trophy – MadeOf – Metal (Trophy / Ganimet)
- Heart – MadeOf – Muscle (Heart / Gönül)
- Jewelery – MadeOf – Silver (Jewelery / Takı)
- Tile – MadeOf – Stone (Tile / Karo)

Table 4.3 shows results after applying Model 5 on a sample of *MadeOf* examples.

Model 6 uses Turkish Wikipedia article abstracts as contexts to score translation candidates. In this case two scoring techniques were applied. The first one (let's refer to this as Model 6.1) compares the number of matching terms between abstracts and the Google Translate definitions. This did correct some of the translations but also resulted in incorrect translations that were previously correct. Some results of this model are listed in Table 4.4.

Table 4.3. Model 5 for MadeOf sample.

INCORRECT	7
CORRECT	83
NO TRANSLATION (Skip)	38
QUESTIONABLE	6
TOTAL	134
ACCURACY	62%

Table 4.4. Model 6.1 for MadeOf sample / matching term count.

INCORRECT	12
CORRECT	101
NO TRANSLATION (Skip)	9
QUESTIONABLE	12
TOTAL	134
ACCURACY	75%

A comparison of Model 5 and Model 6.1 is shown in Table 4.5. The first column shows concepts correctly translated by Model 5. The second and third columns show the results obtained by Model 6.1 for the same concepts. *Questionable* results are translations for a concept that could be accepted but are not possibly the first choice of sense.

Table 4.5. Model 5 vs Model 6.1 for MadeOf sample.

MODEL 5	**MODEL 6.2**	**Result**
Energy - Enerji	Energy – Güç	INCORRECT
Experience - Deneyim	Experience - Olay	INCORRECT
Page - Sayfa	Page - Şövalye eğitimi alan çocuk	INCORRECT
Surface - Yüzey	Surface - Üst	INCORRECT
Wind - Rüzgâr	Wind - Hava	INCORRECT
Pepper - Biber	Pepper - Biber serpmek	INCORRECT
Anchor - Çapa	Anchor - Demir	QUESTIONABLE
Basket - Sepet	Basket – Zembil	QUESTIONABLE
Cult - Tarikat	Cult - Tapınma	QUESTIONABLE
Rock - Kaya	Rock - Taş	QUESTIONABLE
House - Ev	House - Mesken	QUESTIONABLE
Lemonade - Limonata	Lemonade - Limonlu gazoz	QUESTIONABLE
Organism - Organizma	Organism - Canlı varlık	QUESTIONABLE
Tile - Karo	Tile - Kiremit	QUESTIONABLE

The second scoring approach (let's refer to this as Model 6.2) made use of tf-idf scores of terms in abstracts. Google Translate definitions were then used as queries to rank abstracts. The scoring mechanism for Model 6.1 results in counting a lot of irrelevant words that should normally carry no weight if not very small. Repeated use of the semantically less contributing words in abstracts, caused irrelevant articles to score higher. Tf-idf scoring was adopted to remedy the effect of irrelevant word frequencies.

Take the concept “Energy” in the lemma “atom is made of energy”, for example. The selected Google translate definition for “energy” is “the strength and vitality required for sustained physical or mental activity”. This definition contains the words “strength”, “vitality”, “require”, “sustained”, “physical”, “mental” and “activity”. The Turkish Wikipedia article “Güç_(fizik)” has a crosslingual link to the English Wikipedia article “Power_(physics)” and the abstract of this English article contains the words “require” and “physical” in a total of 4 times. These semantically less relevant words cause a higher score for the translation candidate “güç” instead of “enerji”.

Model 6.2 surprisingly performed worse. Table 4.6 lists results for this model.

Table 4.6. Model 6.2 results for MadeOf sample / tf-idf scoring.

INCORRECT	21
CORRECT	94
NO TRANSLATION (Skip)	10
QUESTIONABLE	9
TOTAL	134
ACCURACY	70%

A comparison of Model 5 and Model 6.2 is shown in Table 4.7. The first column shows concepts correctly translated by Model 5. The second and third columns show the results obtained by Model 6.2 for the same concepts.

Despite the number of incorrect translations in general for both scoring methods, Model 6 did improve some results.

- The concept “metal” is now translated as “metal” and not “maden”
- The concept “land” is now translated as “kara” and not “arazi”
- The concept “table” is now translated as “masa” and not “tablo”
- The concept “microchip” is now translated as “mikroçip”

Table 4.7. Model 5 vs Model 6.2 results for MadeOf sample.

MODEL 5	**MODEL 6.2**	**Result**
Energy - Enerji	Energy – Güç	INCORRECT
Cheese - Peynir	Cheese - Peynir kalıbı	INCORRECT
Cloth - Bez	Cloth - Cilt bezi	INCORRECT
Flag - Bayrak	Flag - Flama	INCORRECT
Human - İnsan	Human - İnsani	INCORRECT
Experience - Deneyim	Experience - Olay	INCORRECT
Molecule - Molekül	Molecule - Zerre	INCORRECT
Page - Sayfa	Page - Şövalye eğitimi alan çocuk	INCORRECT
Surface - Yüzey	Surface - Üst	INCORRECT
Pepper - Biber	Pepper - Biber serpmek	INCORRECT
Sink - Lavabo	Sink - Bataklık	INCORRECT
Modelt - Bitki	Modelt - Tesis	INCORRECT
Plastic - Plastik	Plastic - Estetik	INCORRECT
House – Ev	House – Hane	INCORRECT
Brain – Beyin	Brain - Zekâ	QUESTIONABLE
Cult - Tarikat	Cult - Tapınma	QUESTIONABLE
Rock - Kaya	Rock - Taş	QUESTIONABLE
Organism - Organizma	Organism - Canlı varlık	QUESTIONABLE
Salt - Tuz	Salt - Tuzlu	QUESTIONABLE
Sand - Kum	Sand - Kumluk	QUESTIONABLE

There were two cases in the samples where Model 6.1 was actually better than Model 6.2.

- Model 6.2 translated the concept "trophy" as "av hayvanı başı" while Model 6.1 translated as "zafer hatırası" which is the correct sense for "trophy - MadeOf - metal"
- Model 6.2 translated the concept "letter" as "mektup" while Model 6.1 translated as "harf" which is the correct sense for "word - MadeOf - letter"

4.3. Results For Model 7

As the final experiment, Model 7 was applied to the sample *MadeOf* relations. It turned out to produce the most promising results compared to the others, so it was decided to apply this model to all relations.

The following results assume a translation to be correct only if it makes sense in Turkish. Some translations are grammatically not correct, but they capture concepts on both sides of the relation in the correct context. Some of the grammatical errors are caused by tools used and others are results of existing grammatical errors in source examples.

Nearly all relations have examples that don't make much sense in English. There are also many examples which are hard to translate. Some examples are actually asymmetrically divided long sentences. They do not conform to the assumption that there are two concepts on both sides of a relation that on their are isolated units of meaning. These are considered to be incorrect.

Some of these unexpected examples for the MadeOf relation are:

- difference between an entranceway and a patio door: patio door - MadeOf - glass
- pizza usually - MadeOf - tomato sauce, cheese and crust
- stabbing to death may - SymbolOf - dead domination to some person

- graph - MadeOf - set of vertices and a set of edge

Examples above are actually translated into (in respective order):

- giriş kapısı ve veranda kapısı arasındaki fark: veranda kapısı - MadeOf - cam
- pizza genellikle - MadeOf - domates sosu, peynir ve kabuk
- ölüme bıçaklamak - MadeOf - bir kimseye hükmetmek
- grafik - MadeOf - köşe kümesi ve kenar kümesi

It can be said that the first two translations can capture the meaning of their respective source concepts. So generally these examples will be accepted. Accepted translations are somewhat considered to be semantically usable examples in Turkish, although sometimes it might be hard to make sense in proper sentences. There are also grammatically correct translations that semantically do not mean much, but most of these cases were considered to be correct translations too.

Tables 4.8 and 4.9 show examples for *MadeOf* and *SymbolOf* including their translations generated by Model 7.

Relations like *NotHasProperty*, *adjectivePertainsTo*, *adverbPertainsTo* and *NotCapableOf* do not seem to perform well under Model 7. This is because either they contain many hard to translate and sometimes poor quality (unprocessable) examples or are very domain specific and it's hard to translate without a domain specific resource. Adverbs are especially hard to translate into Turkish. Also many of the examples contain multi word phrases or divided sentences.

Some hard to translate and nonsense examples for these relations are:

- accidents can happen to someone who - NotHasProperty - careful
- living at an apartment house you - NotHasProperty - expected to water the grass
- united states president - NotHasProperty - better than other person
- pre jurassic - adjectivePertainsTo - jurassic

Table 4.8. Model 7 results for MadeOf.

Source Example	**Target Example**	**Result**
atom - MadeOf - energy	atom - MadeOf - enerji	Correct
sink - MadeOf - ceramic	lavabo - MadeOf - seramik	Correct
rock - MadeOf - mineral	kaya - MadeOf - mineral	Correct
bike - MadeOf - wheel	bisiklet - MadeOf - tekerlek	Correct
sword - MadeOf - metal	kılıç - MadeOf - metal	Correct
basket - MadeOf - plastic	sepet - MadeOf - plastik	Correct
laptop - MadeOf - chip	dizüstü - MadeOf - yonga	Correct
tree trunk - MadeOf - wood	ağaç gövdesi - MadeOf - ağaçlık	Incorrect
table - MadeOf - tree	tablo - MadeOf - ağaç	Incorrect
tu hermana en bola - MadeOf - tree	tablo - MadeOf - ağaç	Incorrect
nintendo wii - MadeOf - wood	nintendo wii - MadeOf - ahşap	Accepted Correct
blood - MadeOf - serum	kan - MadeOf - serum	Accepted Correct

Table 4.9. Model 7 results for SymbolOf.

Source Example	**Target Example**	**Result**
red color - SymbolOf - blood	kırmızı renk - SymbolOf - kan	Correct
black - SymbolOf - death	siyah - SymbolOf - ölüm	Correct
sword - SymbolOf - combat	kılıç - SymbolOf - mücadele	Incorrect
money - SymbolOf - richness	para - MadeOf - ağırlık (yemek)	Incorrect
pencil - SymbolOf - studying	kalem - SymbolOf - incelemek	Accepted Correct
cloud - SymbolOf - cream cheese	bulut - SymbolOf - yumuşak beyaz peynir	Accepted Correct

- ostial - adjectivePertainsTo - bone
- fenestral - adjectivePertainsTo - fenestra
- unfretted - adjectivePertainsTo - fret
- possessively - adverbPertainsTo - possessive
- conjecturally - adverbPertainsTo - conjecture
- ravishingly - adverbPertainsTo - ravish
- television - NotCapableOf - need to be watered
- duck, but she - NotCapableOf - live
- privacy mean peeping tom's ca - NotCapableOf - watch you

Translations for some examples were accepted although they did not make much sense or grammatically they are not complete, like the following:

- piece of toast / tost parçası - NotHasProperty - alive / canlı
- person / kişi - NotCapableOf - stay the same / aynı kal
- person / kişi - NotCapableOf - taste fax machine / faks makinesini tatmak
- golf ball / golf topu - NotCapableOf - float on water / su üzerinde yüzer

For relations like *mainInterest*, *notableIdea*, *influenced* and *influencedBy* at least one of the concepts is a named entity like a famous scientist, politician or thinker. So either only one side of the relation was translated or the relation was accepted as a whole as the translation. In the cases of *mainInterest* and *notableIdea*, the concept to be translated can be very domain specific, therefore making it hard to translate. Some examples are:

- søren kierkegaard - mainInterest - metaphysics
- walter benjamin - mainInterest- epistemology
- gilles deleuze - notableIdea - schizoanalysis
- martin heidegger - notableIdea - desein

Examples where either concept consists of only stop words were discarded (not translated). Some examples of this type are:

- this - CapableOf - be a bit hard on you
- it - CapableOf - burn the house
- they - CapableOf - become angry
- almost - NotCapableOf - count except in horseshoes
- they - HasA - nice smell

Table 4.10 lists results after applying Model 7 to all relations. For each relation, *Size* is the number of all examples (edges) in that relation for English, *Coverage* is the number of these examples that were processed by Model 7 and *Estimated Accuracy* is the relative frequency of correct translations obtained in randomly selected samples. Relative frequencies of correct translations were calculated analyzing random samples of size 150 for each relation.

Table 4.10. Model 7 results for all relations.

Relation	**Size**	**Coverage**	**Est. Accuracy**
SymbolOf	166	165	84%
DesireOf	280	275	83%
Entails	408	404	79%
NotHasA	409	390	61%
NotIsA	478	402	73%
CreatedBy	503	499	74%
Attribute	639	624	85%
notableIdea	908	908	72%
NotHasProperty	1144	1085	72%
MadeOf	2198	2177	82%
mainInterest	2764	2764	85%

Table 4.10. Model 7 results for all relations (cont.).

Relation	**Size**	**Coverage**	**Est. Accuracy**
adverbPertainsTo	2880	2841	61%
NotCapableOf	2915	2440	60%
HasLastSubevent	3065	3063	66%
adjectivePertainsTo	3313	3297	56%
HasFirstSubevent	4208	4202	62%
NotDesires	4280	4239	71%
Desires	5062	4870	74%
CausesDesire	5176	5158	69%
DefinedAs	6406	6179	61%
HasContext	8851	8615	71%
HasA	9762	9283	62%
ReceivesAction	10429	10090	61%
SimilarTo	11061	10679	74%
MemberOf	12190	12052	55%
PartOf	14151	13791	65%
spokenIn	15590	15427	53%
MotivatedByGoal	15960	15605	68%
Causes	18355	18143	55%
languageFamily	19713	19504	60%
HasProperty	19823	18615	67%
HasPrerequisite	24545	24155	69%
Antonym	26551	24478	71%
field	26732	26450	83%
HasSubevent	26911	26602	62%
knownFor	27519	27224	75%
UsedFor	46522	45381	64%

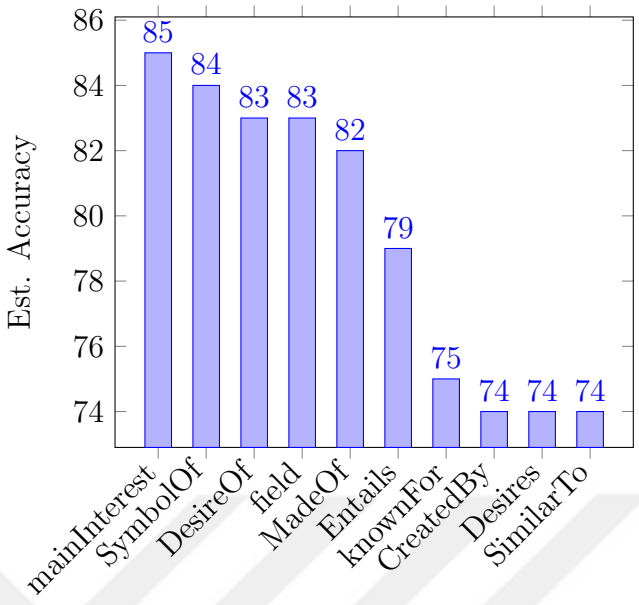


Figure 4.1. Best Performing Relations.

Out of the 58 relations, 37 relations were translated. 10 relations were not included because they were small in size and 9 were not translated because they were huge and the resources of this study were limited. 2 of the relations, namely *influenced* and *influencedBy* were not translated because the majority of nodes for these relations were entity names, specifically famous people in various domains. Figures 4.1 and 4.2 show the best and worst performing relations.

The relative frequencies reflect correct and accepted translations in each randomly selected sample. Accepted translations are translations that are either grammatically slightly incorrect or are seemingly correct but semantically not making sense in a given example.

Table 4.11 shows relative frequencies after grouping the results in order to capture more detail. The *Unprocessable* column refers to source examples that resulted in incorrect translations because they did not make sense or were clearly not usable for translation.

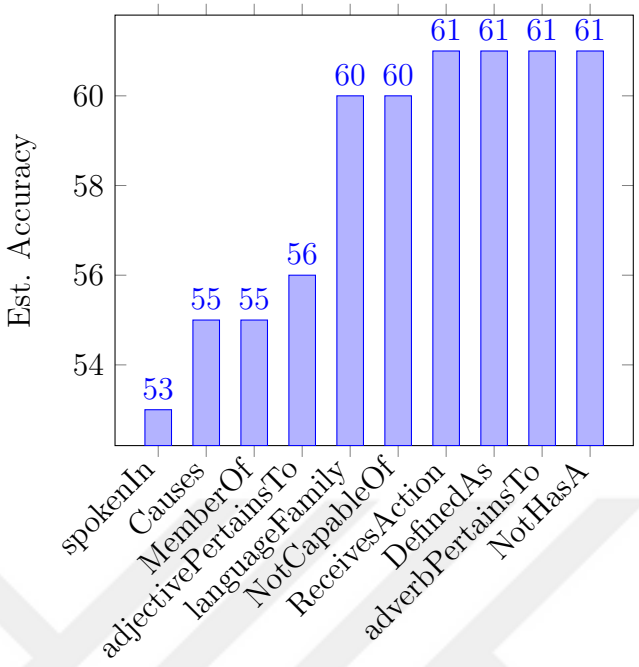


Figure 4.2. Worst Performing Relations.

Table 4.11. Grouped results for all relations.

Relation	**Correct**	**Accepted**	**Incorrect**	**Unprocessable**
SymbolOf	74%	10%	13%	3%
DesireOf	64%	19%	17%	
Entails	70%	9%	21%	
NotHasA	52%	9%	16%	23%
NotIsA	62%	11%	24%	3%
CreatedBy	59%	15%	21%	5%
Attribute	74%	11%	15%	
notableIdea	66%	6%	26%	2%
NotHasProperty	55%	17%	20%	8%
MadeOf	71%	11%	13%	5%
mainInterest	79%	6%	15%	
adverbPertainsTo	48%	13%	37%	2%
NotCapableOf	39%	21%	21%	19%

Table 4.11. Grouped results for all relations (cont.).

Relation	**Correct**	**Accepted**	**Incorrect**	**Unprocessable**
HasLastSubevent	40%	26%	29%	5%
adjectivePertainsTo	48%	8%	43%	1%
HasFirstSubevent	29%	33%	35%	3%
NotDesires	53%	18%	27%	2%
Desires	61%	13%	21%	5%
CausesDesire	55%	14%	31%	
DefinedAs	53%	8%	35%	4%
HasContext	67%	4%	28%	1%
HasA	54%	8%	25%	13%
ReceivesAction	43%	18%	30%	9%
SimilarTo	68%	6%	25%	1%
MemberOf	52%	3%	45%	
PartOf	58%	7%	30%	5%
spokenIn	48%	5%	46%	1%
MotivatedByGoal	51%	17%	32%	3%
Causes	44%	11%	44%	1%
languageFamily	57%	3%	40%	
HasProperty	58%	9%	26%	7%
HasPrerequisite	58%	11%	30%	1%
Antonym	64%	7%	29%	
field	81%	2%	17%	
HasSubevent	49%	13%	38%	
knownFor	70%	5%	24%	1%
UsedFor	47%	17%	35%	1%

Figures 4.3 and 4.4 show the estimated accuracy measurements for each relation against average concept sizes. Start concepts are the concepts positioned to the left hand side of the relation and end concepts are positioned to the right hand side.

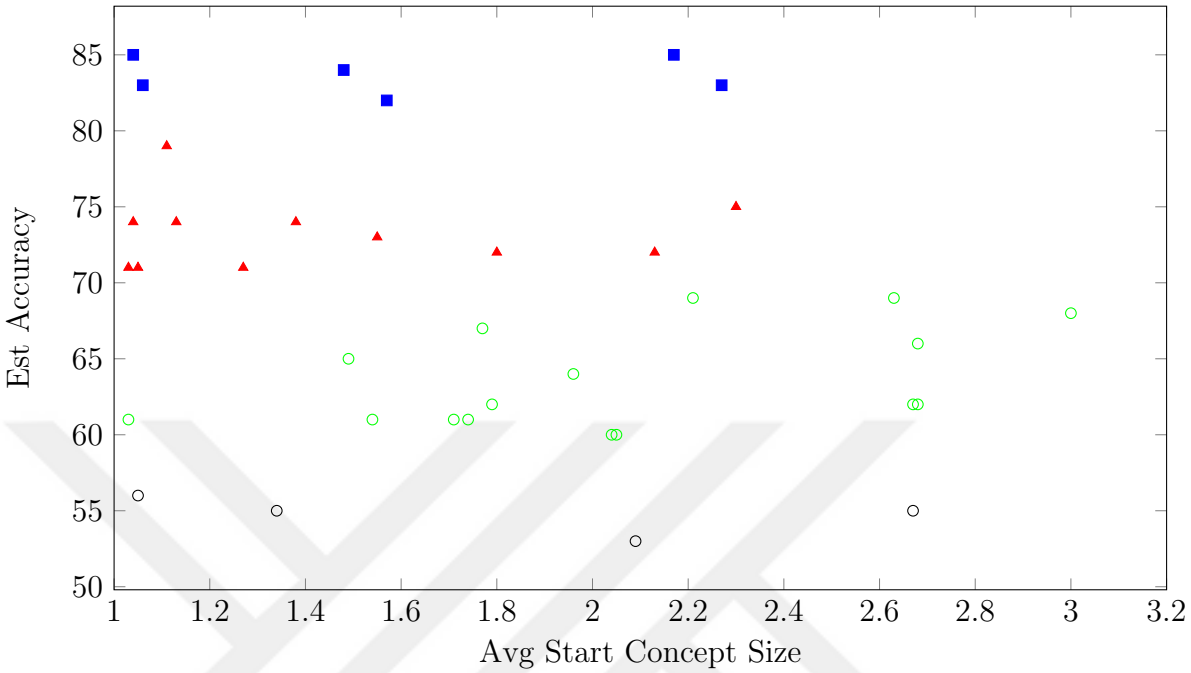


Figure 4.3. Estimated Accuracies vs Average Start Concept Size.

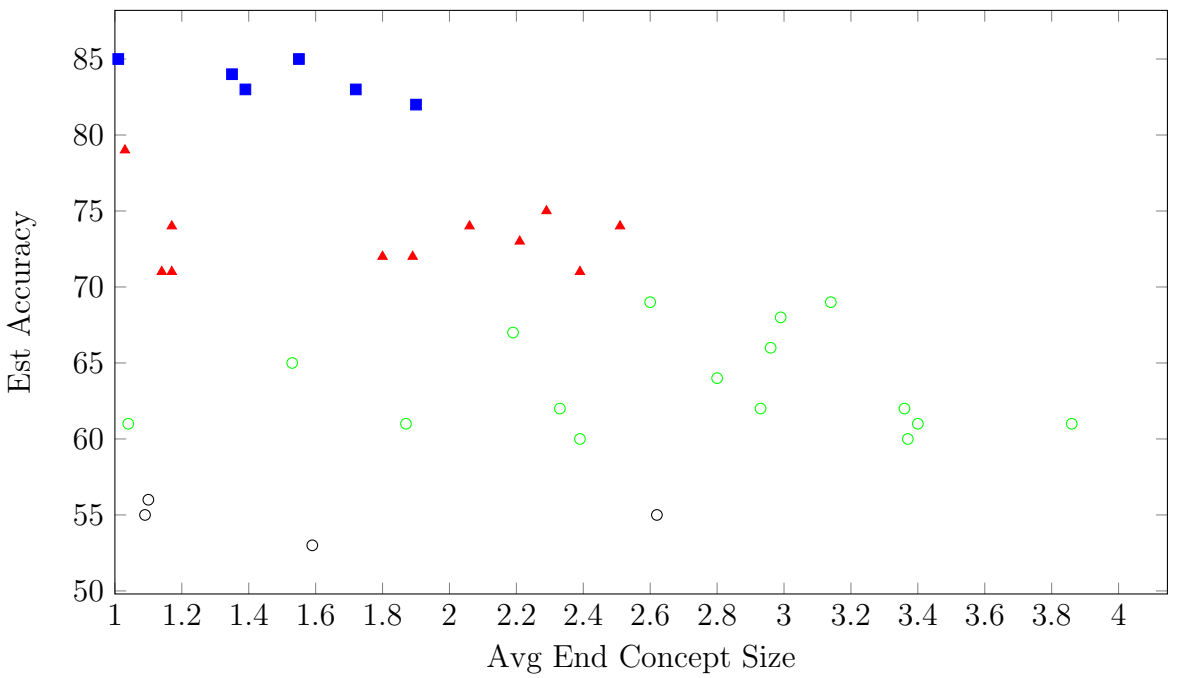


Figure 4.4. Estimated Accuracies vs Average End Concept Size.

Blue squares, red triangles, green circles and black circles represent relations with at least 80% accuracy, between 80% and 70%, between 70% and 60% accuracy and less than 60% accuracy respectively.

The best scoring relations in general contain small and isolated concepts. This is consistent with the assumptions made in Model 7. Relations with relatively larger concepts tend to perform worse.

There are also relations that in general include small concepts but perform badly. These relations contain many low quality examples. This makes them hard to translate and in many cases hard to make sense of. For example, the circles on the very left of each plot are the relations *MemberOf*, *adverbPertainsTo* and *adjectivePertainsTo*. Some examples for MemberOf are:

- genus impatiens - MemberOf - balsaminaceae,
- bomarea - MemberOf - amaryllidaceae,
- ten spined stickleback - MemberOf - gasterosteus,
- gnomic - adjectivePertainsTo - gnome,
- pugilistic - adjectivePertainsTo - box,
- snappishly - adverbPertainsTo - snappish.

Higher scoring relations seem to have a tendency to contain shorter concepts while mid or lower scoring relations are more spread out. This is consistent with the assumption made throughout this study assuming ConceptNet consisted of short concepts on both sides of a relation. Concepts that are longer in size were not considered to be disambiguated and Google Translate or Tureng results were accepted instead.

A combined concept length is the sum of the number of words in each concept. For example, the combined concept length for “waiting for someone - Causes - getting bored” is 5.

Table 4.12 shows how correct and accepted translations are distributed when examples are grouped by combined concept lengths. For instance, among all examples in *MadeOf* having a combined concept length of 4, 65% were accurately translated.

Table 4.12. Estimated accuracies vs concept lengths.

Relation	**Combined Concept Length**				
	2	3	4	5	6
SymbolOf	89%	83%	72%	100%	40%
DesireOf	88%	77%	93%	82%	
Entails	82%	81%	100%		
NotHasA	89%	88%	30%	47%	50%
NotIsA	93%	73%	68%	63%	47%
CreatedBy	89%	71%	83%	44%	40%
Attribute	85%	100%			
notableIdea	100%	75%	76%	54%	71%
NotHasProperty	94%	82%	54%	43%	64%
MadeOf	88%	88%	65%	83%	25%
mainInterest	100%	82%	83%	96%	75%
adverbPertainsTo	62%	100%	33%		
NotCapableOf	89%	80%	56%	77%	48%
HasLastSubevent	78%	42%	81%	73%	57%
adjectivePertainsTo	56%	62%	33%		
HasFirstSubevent		86%	64%	65%	68%
NotDesires	83%	65%	67%	69%	50%
Desires	86%	67%	60%	80%	78%
CausesDesire	100%	71%	60%	65%	63%
DefinedAs	80%	50%	44%	68%	77%
HasContext	69%	76%	75%		100%
HasA	78%	67%	55%	56%	62%

Table 4.12. Estimated accuracies vs concept lengths (cont.).

Relation	**Combined Concept Length**				
	2	3	4	5	6
ReceivesAction	70%	50%	56%	53%	68%
SimilarTo	79%	56%			
MemberOf	55%	56%	33%		100%
PartOf	64%	71%	53%	67%	100%
spokenIn	33%	46%	57%	79%	100%
MotivatedByGoal	44%	83%	79%	68%	61%
Causes	75%	42%	61%	56%	56%
languageFamily		100%	57%	56%	100%
HasProperty	80%	75%	56%	56%	60%
HasPrerequisite	63%	71%	76%	61%	72%
Antonym	71%	64%	71%		
field		81%	83%	90%	100%
HasSubevent	100%	46%	64%	61%	61%
knownFor		77%	80%	66%	95%
UsedFor	57%	76%	61%	52%	63%

5. CONCLUSION

Building resources for Turkish like WordNet, ConceptNet or other common sense knowledge bases is time and resource consuming. Instead, focusing on transfering knowledge from resources in another language seems more feasible and worth looking into.

This thesis study aimed at building a Turkish ConceptNet by translating the existing English ConceptNet. In order to accomplish this, different models were tested. Soon it became apparent that, being able to choose the correct sense of translation for a concept in English was the main challenge to solve.

Word Sense Disambiguation is one of the central issues in computational linguistics. Systems built for domains like text categorization, text summarization, concept extraction, context mining, machine translation have to address the problem of disambiguating senses. Many tools have been proposed or built for languages like English over time. However, progress is hindered due to lack of resources for languages like Turkish.

The work described throughout this thesis attempted to translate as many examples of ConceptNet relations as possible from English into Turkish combining different existing tools. The goal has been to create a network of everyday knowledge for Turkish, a language that lacks a proper common sense knowledge base. The assumption was that ConceptNet would be a good source to translate because it consisted of representations of simple concepts and relations between them.

The method implemented used Google Translate, Tureng, Wikipedia and the Google Search API. Google Translate was used as the primary bilingual dictionary. Wikipedia article extracts were obtained through the Google Search API in order to generate a small, aligned bilingual corpus for each example. Sentences in aligned Wikipedia article extracts were then used to score each translation candidate by match-

ing glosses, examples and synonyms returned by Google Translate. The candidate that obtained the highest score among all article extracts was accepted as the correct translation.

The work described in this thesis, focused on solving WSD for cases where there was very little context. Models were developed and tested incrementally, in respective order, using simple online dictionaries, introducing WordNet and augmenting the Lesk algorithm, using Wikipedia cross lingual links and finally combining Google Translate with Wikipedia article extracts.

Looking at the results, it could be said that the proposed method performed well with some relations having relatively short examples, consisting of isolated and simple nodes. However many relations performed poorly too. This was mainly caused by examples that were either hard to translate into Turkish or had poor quality.

Future work could integrate KeNet [21] and possibly cross lingual WordNet links between Turkish and English into Model 7. There is also the option of improving incorrect translations through feedback implementations or manual corrections.

REFERENCES

1. *Wikipedia*, `https://wikipedia.org/`, accessed in June 2019.

2. Liu, H. and P. Singh, "ConceptNet — A Practical Commonsense Reasoning Tool-Kit", *BT Technology Journal*, Vol. 22, 2004.

3. Singh, P. *et al.*, "Open mind commonsense: knowledge acquisition from the general public", *Conference: On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE, Lecture Notes in Computer Science*, Vol. 2519, pp. 1223–1237, 2002.

4. Chung, H., *GlobalMind - Bridging the Gap Between Different Cultures and Languages with Common-sense Computing*, M.Sc. Thesis, Massachusetts Institute of Technology, 2006.

5. *Wiktionary*, `https://en.wiktionary.org`, accessed in June 2019.

6. *WordNet*, `https://wordnet.princeton.edu/`, accessed in June 2019.

7. *DBPedia*, `https://dbpedia.org`, accessed in June 2019.

8. *Umbel*, `http://umbel.org`, accessed in June 2019.

9. *OpenCyc*, `https://www.cyc.com/opencyc/`, accessed in June 2019.

10. *ConceptNet5*, `http://conceptnet5.media.mit.edu/`, accessed in June 2019.

11. Speer, R. and C. Havasi, "Representing General Relational Knowledge in ConceptNet 5", *Language Resources and Evaluation*, 2012.

12. Speer, R., *ConceptNet5 Languages*, 2016, `https://github.com/commonsense/conceptnet5/wiki`, accessed in June 2019.

13. Anacleto, J. C. *et al.*, "How Can Common Sense Support Instructors with Distance Education?", *Workshop em Informática na Educação (sbie) 2006 XVII Simpósio Brasileiro de Informática na Educação - SBIE - UNB/UCB*, pp. 328–337, 2006.

14. *Induction of Linguistic Knowledge Research Group*, `https://ilk.uvt.nl/`, accessed in June 2019.

15. Eckhardt, N., *A Kid's Open Mind Common Sense*, M.A. Thesis, Tilburg University, 2008.

16. Stamou, S. *et al.*, "BALKANET: A Multilingual Semantic Network for Balkan Languages", *First International Wordnet Conference, Mysore, India*, pp. 12–14, 2002.

17. Cristeau, D. *et al.*, "BalkaNet: Aims, Methods, Results and Perspectives. A General Overview", *Romanian Journal of Information Science and Technology*, Vol. 7, pp. 9–43, 2004.

18. Fellbaum, C., "WordNet and wordnets", *Encyclopedia of Language and Linguistics. Elsevier.*, Vol. 3, pp. 665–670, 2006.

19. Oflazer, K. *et al.*, "Building a Wordnet for Turkish", *Romanian Journal of Information Science and Technology*, Vol. 7, pp. 163–172, 2004.

20. *KeNet*, `http://haydut.isikun.edu.tr/wordnet.ui-1.0/`, accessed in June 2019.

21. Yildiz, O. T. *et al.*, "Constructing a WordNet for Turkish Using Manual and Automatic Annotation", *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 17, No. 3, Article 24, 2018.

22. Oflazer, K. *et al.*, "SentiTurkNet: a Turkish polarity lexicon for sentiment analysis", *Language Resources and Evaluation (LREC)*, Vol. 50, pp. 667–685, 2016.

23. Ozcan, S. and M. F. Amasyali, "Turkish Commonsense Database (CSDB) And Csoyun (A Game With A Purpose)", *Sigma Journal of Engineering and Natural Sciences*, Vol. 32, pp. 116–127, 2014.

24. Yazici, E. and M. F. Amasyali, "Automatic Extraction of Semantic Relationships Using Turkish Dictionary Definitions", *EMO Bilimsel Dergi*, Vol. 1, pp. 1–13, 2011.

25. *Google Translate*, `https://translate.google.com/`, accessed in June 2019.

26. Montazery, M. and H. Faili, "Unsupervised Learning for Persian WordNet Construction", *Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria*, pp. 302–308, September 2011.

27. Montazery, M. and H. Faili, "Automatic Persian WordNet Construction", *23rd International conference on computational linguistics, Beijing, China*, pp. 846–850, 2010.

28. Sarrafzadeh, N. *et al.*, "Cross Lingual Word Sense Disambiguation for Languages with Scarce Resources", *24th Canadian Conference on Artificial Intelligence*, pp. 347–358, 2011.

29. Sivakumar, J. and A. Anthoniraj, "Automatic Word Sense Disambiguation Using Wikipedia Link Structure", *International Journal of Computer Science Communication Networks*, Vol. 3, pp. 147–151, 2013.

30. Lesk, M., "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone", *Proceedings of SIGDOC-86: 5th International Conference on Systems Documentation*, pp. 24–26, 1986.

31. Gaillard, B. *et al.*, "Query Translation using Wikipedia-based resources for analysis and disambiguation", *EAMT 2010 - 14th Annual Conference of the European Association for Machine Translation*, 2010.

32. Agirre, E. and G. Rigau, "A Proposal for Word Sense Disambiguation using Conceptual Distance", *International Conference on Recent Advances in Natural Language Processing. Tzigov Chark, Bulgaria.*, 1995.

33. Speer, R. *et al.*, "Coarse Word-Sense Disambiguation Using Common Sense", *AAAI Fall Symposium: Commonsense Knowledge*, 2010.

34. Manning, C. *et al.*, "The Stanford CoreNLP Natural Language Processing Toolkit", *Proceedings of 52Nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

35. Akın, A. A., *Zemberek-NLP*, 2019, `https://github.com/ahmetaa/zemberek-nlp`, accessed in June 2019.

36. *Wordreference*, `https://www.wordreference.com/`, accessed in June 2019.

37. *Tureng - Multilingual Dictionary*, `https://tureng.com/tr/turkce-ingilizce`, accessed in June 2019.

38. *Wikimedia Foundation*, `https://wikimediafoundation.org/`, accessed in June 2019.

39. Yildiz, E. *et al.*, "The Effect Of Parallel Corpus Quality vs Size In English-To-Turkish SMT", *Fourth International Conference on Computer Science, Engineering and Applications (ICCSEA-2014)*, pp. 21–30, 2014.

40. *Turkish Wikipedia*, `https://tr.wikipedia.org`, accessed in June 2019.

41. *Google Custom Search*, `https://developers.google.com/custom-search/`, accessed in June 2019.

42. Celik, K., *A Comprehensive Analysis Of Using Wordnet, Part-Of-Speech Tagging, And Word Sense Disambiguation In Text Categorization*, M.Sc. Thesis, Bogazici University, 2009.

43. Banarjee, S., *Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet*, M.Sc. Thesis, University of Minnesota, 2002.