Annotation Guidelines for Labeling English-Dutch Cognate Pairs

Version 1.0

LT3 Technical Report – LT3 19-02

Sofie Labat, Lore Vandevoorde and Els Lefever

LT3 – Language and Translation Technology Team
Department of Translation, Interpreting and Communication
Ghent University

August 19, 2019

Contents

Chapter 1

Introduction

Processing semantic information is a complex task, especially due to the polysemous nature of language. This is no different for the task of cognate detection. In general linguistics, the term *cognate* is defined as a "language or a linguistic form which is historically derived from the same source as another language/form" (Crystal, 2008, page 83). The assumption of common etymology is, however, often disregarded in the literature, because certain research areas such as psycholinguistics or NLP tend to shift their focus from diachronic to perceptual relatedness (Shlesinger and Malkiel, 2005; Mitkov et al., 2007; Schepens et al., 2013; Hansen-Schirra, Nitzke, and Oster, 2017). We follow this second strand of research in that we define cognates as words with high formal and semantic cross-lingual similarity.

The ability to distinguish cognates from non-cognates[1] (and especially) false friends is an important skill for second language learners. Similarly, source language interference is a problem often experienced by translators that is partly caused by the influence of cognates and false friends. Research in natural language processing can address these bottlenecks by for instance developing computer tools that aid second language users. Unfortunately, extensive lists of known cognates and false friends which could be used for the development of such systems are hard to find and expensive to compose, since they require a considerable amount of time and effort from trained lexicographers (Schepens, Dijkstra, and Grootjen, 2012). Especially for low resource languages, this constitutes a serious issue.

The following guidelines propose a clearly defined method for the automatic generation of English-Dutch bilingual term lists of candidate cognates and false friends. In a next step, a human annotator has to manually assign classification labels to the candidate pairs. The resulting data frame or gold standard is context-independent. Hence, it can be used for both the development and the evaluation of machine learning models that deal with cognate detection. Besides its applications in natural language processing, the gold standard can also form an important new resource for further research in e.g. translation studies or psycho-linguistics.

[1] In psycholinguistics, non-cognates are "L1-L2 translation pairs that share meaning but not form" (de Groot, 2011, page 121). We deviate from this definition in that we define non-cognate word pairs as words which are simply no cognates. Therefore, false friends and alignment errors are also classified as non-cognates.

Chapter 2

Data Preparation: Generation of Candidate Cognate Pairs

To select a list of candidate cognate pairs, unsupervised statistical word alignment using GIZA++ (Och and Ney, 2003) was applied on the Dutch Parallel Corpus (DPC). This high-quality parallel corpus for Dutch, French and English consists of more than ten million words and is sentence-aligned. It contains five different text types and is balanced with respect to text type and translation direction. The automatic word alignment on the English-Dutch part of the DPC resulted in a list containing more than 500,000 translation equivalents. A first selection was performed by applying the Normalized Levenshtein Distance (NLD) (as implemented by Gries (2004)) on this list of translation equivalents and only considering equivalents with a distance greater than or equal to 0.5. This resulted in a list with 28,503 Dutch-English candidate cognate pairs, which was manually labeled.

Chapter 3

Annotation Scheme

3.1 General Overview

To create the gold standard for cognate detection, an extensive set of guidelines was established. The following section considers these guidelines in more detail. Firstly, twelve clearly defined labels were introduced to categorize the candidate cognate pairs:

1. **Cognate**: words which have a similar form and meaning in all contexts. Conform with our working definition for cognates, the source and target words do not need to be etymologically related (e.g. *organisation - organisatie*).
2. **Cognate agreement**: words which, like the *Cognate* labeled pairs, have a similar meaning and form. They differ from *Cognate* pairs in that there is no agreement between the members of the pair (e.g. *organisations - organisatie*).
3. **Cognate PoS**: although words labeled under this category also have a similar meaning and form, they disagree on their PoS-tag (e.g *winter - winterse*).
4. **Partial cognate**: words which have a similar form, but only share the same meaning in some contexts (e.g. EN: *argument* meaning either (i) dispute or (ii) fact/ proof in English; NL: *argument* meaning fact/ proof in Dutch).
5. **False friend**: words which have a similar form, but which have a different meaning in all contexts (e.g. *brave - braaf*). It is important to notice that this label should only be used when we are dealing with word pairs that solely occur as false friends and never as a cognates. If the members of the pair can also occur as cognates, we should label the pair as "Partial cognate".
6. **Neologism**: A word pair of which one of the members is a newly created word that does not yet occur in the dictionary, but the word is also likely to be used in that particular language and often it already occurs in colloquial speech (e.g. *abstentionism - abstentionisme; gangsta - gangsta*).
7. **Proper name**: proper nouns (e.g. persons, companies, cities, countries, etc. and their derivations (e.g. *American - Amerika*). It is important to note that a successful dictionary look-up never overrules this "Proper name" annotation.
8. **Number**: words pairs of which either one or both members contain a number (e.g. *adm12006e*) or a special character (cf. @, , #, /, , +). This label also includes Roman numerals (e.g. *VI*).

9. **No standard**: words that do not occur in the dictionary (e.g. *num_connectors*).

10. **Other word class**: words that do not belong to the four standard word classes (i.e. verbs, nouns, adjectives, and adverbs) in which we are interested (e.g. interjections: *well - welnu*; pronouns: *himself - hijzelf*).

11. **Error**: word alignment errors and compound nouns of which one part is a cognate, but the other part is missing in one of the languages[1] (e.g. *other - achter; peripherals - aansturingsperipherals*).

12. **Error proper name**: word alignment errors in which at least one of the members of a word pair is a proper noun (e.g. *series - stevie*).

To decide on the correct label, we adopted a context-independent approach that looks at the meaning potential of words instead of considering their different context-specific meanings. We applied the following procedure:

(i) For every candidate cognate pair, the dictionary Grote Van Dale[2] (henceforth: VD) was consulted;

(ii) The English word is looked up in the VD, e.g. *salon*;

(iii) The Dutch translation is inspected in the VD, e.g. *salon: "nice room"* and *salon: (room for) gathering of people (e.g. from the literary world)*;

(iv) In case of doubt, additional dictionaries as the *OED* (*Oxford English Dictionary*), *The Free Online Dictionary*, and *MWB* (*Mijnwoordenboek*)[3] are consulted.

Based on the previously obtained information, a decision is made and one of the twelve labels is assigned to the candidate cognate pair. In a somewhat less fine-grained overview, we summarize that in case all meanings of the Dutch word correspond with the English word, they receive the "Cognate" label. If only part of the Dutch meanings correspond with the English word, that pair should be annotated as "Partial cognate". In case the words have different meanings in all possible contexts, they are annotated as "False friend[s]". Finally, when the candidate cognate pairs appear to be the result of an alignment error, we label these pairs as "Error". If one is in doubt about whether the words are full or partial cognates, it is best to consistently assign the more general label (i.e. "Partial cognate") to the word pair. The same reasoning applies to full and partial false friends.

In the previous example with *salon*, the word pair was annotated as "Cognate", since the meaning potential of the English *salon* corresponded in all senses to the meaning potential of the Dutch *salon*. An example of partial cognates are *agent-agent*: The Dutch *agent* refers both to (i) a police man and to (ii) a representative (e.g. business representative). As only the second meaning of the Dutch word is expressed by the English *agent*, these words are considered partial cognates. More fine-grained distinctions were actually made while constructing the annotation guidelines. These have been recorded in great detail and they can be found in Section 3.2.

[1] This second case highlights one of the shortcomings of our current study. Since we are only looking at single words, Dutch compound nouns are quite often taken into account as these are usually written as one word. Conversely, not all English compound nouns are considered, due to the fact that the nouns which compose the compound noun are often separated by white spaces. In future research, one might extend the current gold standard by also considering this type of n-grams.

[2] https://www.vandale.be

[3] https://www.oed.com, https://www.thefreedictionary.com, and https://www.mijnwoordenboek.nl, resp.

3.2 Detailed Annotation Guidelines

The following section discusses the somewhat more technical, fine-grained decisions that were taken while creating the gold standard of English-Dutch cognate pairs. It is subdivided in four sections: the first section deals with lexical issues; the second section deals with semantic issues; the third section explains what was done in case label conflicts occurred; the fourth and final section introduces some decisions made regarding inflected words.

3.2.1 Lexical Issues

1. **Abbreviations:** Abbreviations are included in the annotation system when they occur in the VD (e.g. *aids, id*). If they do not occur in the VD, they are annotated as "No standard".

2. **Chemical and medical compound nouns:** They are incorporated in the annotation system when they either completely occur in the VD (e.g. *phenobarbital - fenobarbital*), or when the constituents which form the compound noun occur in the VD (e.g. *bromobutyl - bromobutyl*; *polyglutamate polyglutamaat*).

3. **Festivities:** Festivities are labeled as "Proper noun[s]" (e.g. *Christmas - Kerstmis; Sukkot - Soekot*).

4. **Greek/ Latin loan words (e.g. medical terms):** If these words are correct and accepted in both languages, even though they do not occur in the VD, they will be classified as cognates (e.g. *accumbens*). One exception is made as these words are not incorporated in our system when they refer to a specific species. In that case, the word pair is labeled as "Proper noun" (e.g. *falciparum*: a word which is used in combination with Plasmodium to refer to a species of unicellular parasites).

5. **Medical conditions:** Medical conditions are always included in the system, even if they do not occur in the VD (e.g. *hypokalaemia - hypokalimie*).

6. **Religious words:** The only proper names that are included in the current gold standard are words that refer to religious things / concepts (e.g. *islam - islam*).

7. **Substance names:** They are included in the annotation system, unless they do not occur in the VD. In the latter case, they are annotated as "Proper name" (for types of medication) or as "No standard" (for chemical substances), e.g. the pair *zyprexa - zyprexa* is excluded.

8. **Switched EN-NL words:** If the Dutch word is listed in the column with English words and the English word is similarly placed in the column of Dutch words, then the two words are switched if the reverse combination (i.e. EN word in EN column and NL word in NL column) does not yet occur in the data frame.

9. **Words containing a hyphen:** Words as *marketing-* which contain a hyphen are left out of the annotation system when the same word without hyphen (cf. *marketing*) also occurs in the data frame. When such words are left out, they are annotated with the label "No standard".

3.2.2 Semantic Issues

1. **Colloquial and archaic senses:** If a word has a whole range of word senses, we will not take into account very archaic and colloquial (cf. informal, slang) word senses in order to make a decision on the classification label.

2. **Verbs requiring a different number of arguments:** A candidate cognate pair can consist of verbs which, depending on the language and context, require a different number of objects or arguments (cf. intransitive and transitive verbs). We decided to not make any distinction between the verbs, if their forms and meanings correspond largely (e.g. *lunch - lunch; sleep* (cf. *the hotel sleeps 80 people*) - *slapen*).

3.2.3 Label Conflicts

1. **Different PoS and agreement:** When the PoS between the source and target words differs and there is no agreement between the two words, we will opt for the "Cognate PoS" label and write "agreement" in the comment column. An exception is made for nouns that only occur in plural forms (e.g. *mechanics*). An example is the word pair *administrative - administrations* which is labeled as "Cognate PoS" and has "agreement" standing in its comment column. It is important to notice that the pair *administrative - administration* is annotated as "Cognate PoS" without any extra comments.

2. **Precedence in labels:** When several labels or classifications are applicable, the following order applies: Cognate PoS (1) > Cognate agreement (2) > Partial cognate (3). For example, *advise - advies* is annotated as "Cognate PoS", while the pair *advice - advies* is labeled as "Partial cognate". Similarly, the pair *affairs - affaire* is an instance of "Cognate agreement", while the pair *affair - affaire* is classified as "Partial cognate".

3. **Word form which can be labelled with different PoS-tags:** In this case, there are two options. For the first option, we will only look at the meaning potential of the most likely PoS-tags for the candidate cognate pair, after firstly having considered the PoS-tags of the individual instances for that pair. This method is applied when the meanings of a word under different PoS-taggings are still closely related. For example, *swarm* can be a verb and a noun, but since it is translated as *zwerm* in Dutch, we will only look at the meanings of the noun *swarm*. The other option is that we will consider all the meanings of that word, linked to different PoS-tags. We adopt this second approach when the meanings of a word under different PoS-taggings are not so closely related.

4. **"No standard" and "Error":** If both "No standard" and "Error" are correct labels for a word pair, then we label the pair as "No standard".

3.2.4 Inflections and Derivations

1. **Adjectives derived from proper names:** Adjectives referring to proper names as persons (e.g. *Napoleonic - Napoleonistisch*) and geographic locations/nationalities (e.g. *Neapolitan - Napolitaans, Italian - Italiaans*) are all excluded and annotated as "Proper name".

2. **Comparative and superlative:** Uninflected adjectives and adjectives that are inflected to comparative or superlative forms differ in their agreement. Hence, the members of a cognate pair with differing inflections are annotated as "Cognate agreement".

3. **Diminutives:** In some languages, diminutives are used more often than in others (cf. NLdim > ENdim). When the meaning potential of a diminutive source word coincides with the meaning potential of the target word in the other language, the label "Cognate" applies. When this is partially the case, the pair is classified as "Partial cognate" (e.g. *addresses - adresjes*). Finally, the pair is annotated as "Error" when the meaning potentials of both words never coincide with each other.

4. **Dutch adjectives ending in *-e*:** No distinction is made between the standard form and the infected form of the Dutch adjective (e.g. *ideal - ideaal* and *ideal - ideale* are labeled in the same way).

5. **Gender:** Two words of opposite gender differ in their agreement (e.g. *actresses - acteurs* is labeled as "Cognate agreement").

6. **Infinitive and present participles:** For the English verbs, we decided to treat the infinitives and present participles as being the same form. This decision is motivated by the fact that the English present participle is frequently translated by an infinitive in Dutch. A result of this decision is that candidate cognate pairs as *combining - combineren* have the same agreement.

7. **Subjunctive mood:** Dutch words which are conjugated in this verb mood are annotated as "No standard", since the form is rather archaic and no longer used. An exception can be made in case a verb in subjunctive mood appears in a fixed expression.

References

Crystal, D. 2008. *A Dictionary of Linguistics and Phonetics*. The language library. Blackwell, 6th edition.

de Groot, A. M. B. 2011. *Lanugage and Cognition in Bilinguals and Multilinguals: An Introduction*. Taylor & Francis.

Gries, S. T. 2004. Shouldn't It Be Breakfunch? A Quantitative Analysis of Blend Structure in English. *Linguistics*, 42(3):639–667.

Hansen-Schirra, S., J. Nitzke, and K. Oster. 2017. Predicting cognate translation. In S. Hansen-Schirra, O. Czulo, and S. Hofmann, editors, *Empirical modelling of translation and interpreting*. Language Science Press, chapter 1, pages 3–22.

Mitkov, R., V. Pekar, D. Blagoev, and A. Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, 21(1):29–53.

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

Schepens, J., T. Dijkstra, and F. Grootjen. 2012. Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, 15(1):157–166.

Schepens, J., K. Paterson, T. Dijkstra, F. Grootjen, and W. J. B. van Heuven. 2013. Cross-Language Distributions of High Frequency and Phonetically Similar Cognates. *PLOS ONE*, 8:1–15.

Shlesinger, M. and B. Malkiel. 2005. Comparing Modalities: Cognates as a Case in Point. *Across Languages and Cultures*, 6(2):173–193.