The University of Oxford
MSc (Mathematics and Foundations of Computer Science)

Applying Distributional Compositional Categorical Models of Meaning to Language Translation

Brian Tyrrell

Abstract. The aim of this essay is threefold: first we will use vector space distributional compositional categorical models of meaning to compare the meaning of sentences in Irish and in English (and thus ascertain when a sentence is the translation of another sentence). Then we shall build an algorithm which translates nouns by understanding their context, using a conceptual space model of cognition. Finally we briefly introduce metrics on **ConvexRel** and use them to determine the distance between concepts (and determine when a noun is the translation of another noun). Although these methods can be applied to many other languages, this essay will focus on applications to Irish.

Contents

> Luke: "Do you understand anything they're saying?"
> C-3PO: "Oh, yes, Master Luke! Remember that I am fluent in over six million forms of com-"
> Han: "What are you telling them?"
> C-3PO: "Hello, I think. I could be mistaken."
>
> \- *Star Wars: Episode VI - Return of the Jedi*

Date: May 17, 2018.

1. Introduction

The raison d'être of Distributional Compositional Categorical (henceforth referred to as DisCoCat) Models of Meaning originates in the oft quoted mantra of the field:

> *"You shall know a word by the company it keeps."*
> -John R. Firth, *A synopsis of linguistic theory 1930-1955*, (1957).

The broad idea of such models in natural language processing is to marry the semantic information of words with the syntactic structure of a sentence using category theory to produce the whole meaning of the sentence. The semantic information of a word is captured (in early models [6, 14, 15]) by a vector in a tensor product of vector spaces using a corpus of text to represent a given word in terms of a fixed basis of other words; i.e. by distributing the meaning of the word across the corpus. In later models ([3]) convex spaces are used instead of vector spaces in an effort to capture the representation of words in the human mind. In simpler terms it is the context of a word rather than the word itself which gives meaning, so the older words of Shakespeare still guide our hands:

> *"That which we call a rose, by any other word would smell as sweet."*
> -William Shakespeare, *Romeo and Juliet.*

It is the focus of this essay to exploit the existing DisCoCat structure in two directions. First, we shall use a vector space model of meaning, defined by Coecke et al. [6] and introduced in *Section 3*, to assign meaning to sentences in English and then in Irish. These meanings are then compared via an inner product on the shared sentence space of English and Irish vector space models of meaning in *Section 4*. We discuss the results of this on a complicated sentence in *Section 4.1*.

Before this, we must determine the Lambek pregroup grammar structure for Irish (which does not exist in the current literature) and, as we shall see in *Section 2*, is nontrivial in some aspects. The ideas presented here and in the subsequent sections can be applied to many other languages, however the author has chosen Irish due to its relative rareness in literature and its high regularity and uniformity in grammar and verb structure. For instance, across all of Irish there exist exactly eleven irregular verbs; with the exception of these eleven, every other verb can be conjugated in an extremely efficient and easy manner. To aid the reader with a language they may not be familiar with, all Irish words are presented coloured green.

After thoroughly discussing Irish and English vector space models of meaning, we will extend this treatment to *conceptual space* models of meaning in the category **ConvexRel** (defined by Bolt et al. [3] using the work of Gärdenfors [11]). *Sections 5 & 6* detail a novel solution towards the generation of conceptual spaces algorithmically from given corpora, and *Sections 6.1 & 6.2* test this solution. At the time of writing there does not exist an algorithmic approach to generating a conceptual space for any given noun that the author is aware of. The results of these sections allow us to use the conceptual space model of meaning to translate nouns in Irish to English based on the context of the noun in Irish, which we preform in *Section 7* using metrics created from the theory of Marsden and Genovese [24].

We begin the story by determining the Lambek pregroup grammar structure for Irish.

2. Lambek Pregroup Grammar Structure for Irish

It is mentioned throughout the literature of the subject, but in particular by Coecke et al. [6] and Grefenstette and Sadrzadeh [15], that DisCoCat models reconcile two aspects of natural language:

Meaning: Vector spaces (or, later on, convex 'conceptual' spaces) can be used to assign meanings to words in a language.

Grammar: Pregroups (an introduction given in [6], a more detailed discussion in relation to grammar in [20]) are used to assign grammatical structure to sentences.

On that second point, it is the diagram of a reduction in a pregroup that produces the 'from-meaning-of-words-to-meaning-of-a-sentence' map which gives a sentence a concrete, comparable meaning based on its contents and grammatical structure. Consider the following example:

Example 2.1. Lambek [20] has a more detailed approach to language than what we need in order to build an operational DisCoCat model; in particular, he considers six basic types (subject, third person singular subject, declarative sentence in present tense, ... etc.) and hand-constructs type assignments of linguistic structures as more complicated grammatical phenomena are encountered. For our purposes we shall consider the simpler pregroup grammar discussed in [6], where the basic types are nouns (n), declarative statements (s), infinitives of verbs (j) and glueing types (σ). Common grammatical structures, such as the following, are assigned compound types:

1. **Adjectives** are assigned the type nn^l,
2. **Transitive verbs** are assigned the type $n^r s n^l$,
3. **Adverbs** are assigned the type $s^r s$.

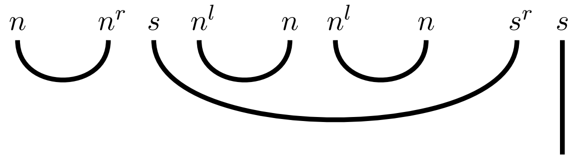
So the example sentence

Colin flys green aeroplanes expertly

has the type assignment

$$n \quad n^r s n^l \quad n n^l \quad n \quad s^r s$$

which has a reduction diagram:



which yields a map:

$$f := (\epsilon_S^r \otimes 1_S) \circ (\epsilon_N^r \otimes 1_S \otimes \epsilon_N^l \otimes \epsilon_N \otimes 1_S \otimes 1_S),$$

where

$$f : N \otimes (N \otimes S \otimes N) \otimes (N \otimes N) \otimes N \otimes (S \otimes S) \to S$$

and N, and S are vector spaces corresponding to nouns and sentences respectively[1].

The map f is in fact a morphism of the compact closed category **FVect** $\times P$, where P is the free pregroup generated by the four basic types above, realised as a compact closed category. The meaning of the sentence "Colin flys green aeroplanes expertly" can be realised completely in S due to f:

[1]In the case of S, the *meaning* of the sentence is a vector in S and hence S is known as the *sentence space*.

$$\begin{aligned}
&\overrightarrow{\text{Colin flys green aeroplanes expertly}}\\
&= f(\overrightarrow{\text{Colin}} \otimes \overrightarrow{\text{flys}} \otimes \overrightarrow{\text{green}} \otimes \overrightarrow{\text{aeroplanes}} \otimes \overrightarrow{\text{expertly}})\\
&= \sum_{ijk,m,p} c_{ijk}^{\text{flys}} d_{km}^{\text{green}} e_{jp}^{\text{expertly}} \langle \overrightarrow{Colin} | \overrightarrow{n}_i \rangle \langle \overrightarrow{aeroplanes} | \overrightarrow{n}_m \rangle \overrightarrow{s}_p
\end{aligned}$$

(where we are working under the assumption the bases of N and S are orthonormal.) ♢

Of course, this is all built exclusively through English, but there are no barriers to moving to a different language; Lambek et al. [1, 2, 5, 21] detail a pregroup structure for French, Arabic, Latin and German, respectively. However the author cannot find evidence of the same treatment in Irish. Thus, in order to create Irish DisCoCat models, we must create a 'Lambek Pregroup Grammar' for the language.

2.1. **Irish Grammatical Structure.** For our purposes, we do not need a structure as complicated as Lambek's work [20], rather we shall mirror the English approach; four basic types - nouns (n), declarative statements (s), infinitives of verbs (j) and glueing types (σ). We hand construct the following compound types:

(1) **Transitive verbs** are assigned the type $sn_2^l n_1^l$, where n_1 is the type of the subject and n_2 is the type of the object. This is because Irish follows the rule *Verb Subject Object*. The only exception to this is the copula is, which we assign the type $sn_1^l n_2^l$ - this verb is used in declarative sentences that are absolutely true.

For example, even though the Irish for the verb "to be" is bí, which in the present tense is tá, one would say

Is dochtúir mé	for "I am a doctor" and
Tá scamaill sa spéir	for "There are clouds in the sky",

(as that second sentence is time and location dependant, thus is not absolutely true).

The reason we include indices in our type assignments in Irish is for clarity only: Irish sentences are not linear in their grammar, unlike English, thus we must keep track of words more carefully.

(2) **Adjectives** are assigned the type $n^r n$, where n is the type of the noun the adjective is describing. This is because Irish follows the rule *Noun Adjective*.

(3) **Adverbs** are assigned the type $s^r s$; they appear at the end of sentences.

(4) **Prepositions** as *whole phrases* are assigned the type $n^r n$. This is because Irish follows the rule *Preposition Noun*, as in English, so we give the same type assignment as in [14]. Note that prepositions in Irish always come before the noun, and adjectives after, so we cannot confuse them.

It should be noted that Irish (sometimes) modifies the noun after a preposition directly by inserting an urú or a séimhiú into the noun - additional letters to change the sound of the word. So, for example, whilst *table* is bord, *on the table* becomes ar an **m**bord. This is a sign that correlates with the change in type assignment of the affected noun.

Let us give some examples to demonstrate.

Example 2.2.

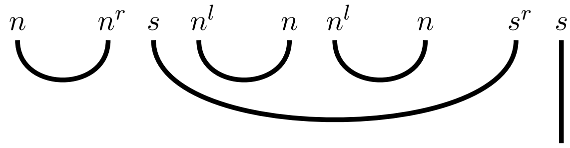
(1) English sentence: "I got a red jumper yesterday". In Irish: "Fuair mé geansaí nua inné".

I got a new jumper yesterday

has the type assignment

$$n \quad n^r s n^l \quad n n^l \quad n \quad s^r s$$

which has a reduction diagram:



In Irish,

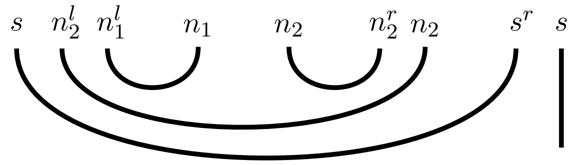
Fuair mé geansaí nua inné

Got I jumper new yesterday

has the type assignment

$$s n_2^l n_1^l \quad n_1 \quad n_2 \quad n_2^r n_2 \quad s^r s$$

which has a reduction diagram



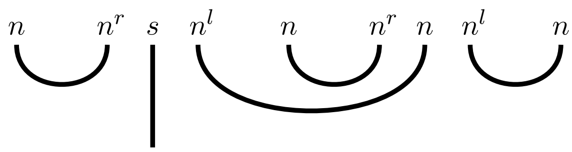
(2) English sentence: "She cooked a plate of tasty sausages." In Irish: "Cócaigh sí pláta ispíní blasta". Note that in Irish the preposition 'of' does not physically appear in the sentence; grammatically, however, it is still present.

She cooked a plate of tasty sausages

has the type assignment[2]

$$n \quad n^r s n^l \quad n \quad n^r n n^l \quad n$$

which has a reduction diagram:



On the other hand, in Irish:

[2]Note that "of tasty" has the assignment $n^r n n^l$. It is specified in [14] that the whole prepositional phrase should be given type $n^r n$; here the phrase is "plate of tasty sausages" so is given type $n^r(nn^l)n = (n^r nn^l)n$.

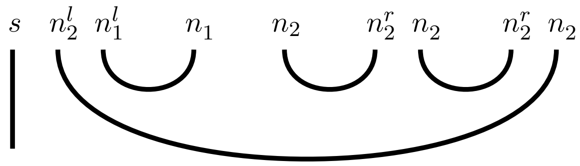
Cócaigh sí pláta ispíní blasta

Cook she plate (of) sausages tasty

has the type assignment[3]

$$sn_2^l n_1^l \quad n_1 \quad n_2 \quad n_2^r n_2 \quad n_2^r n_2$$

which has a reduction diagram:



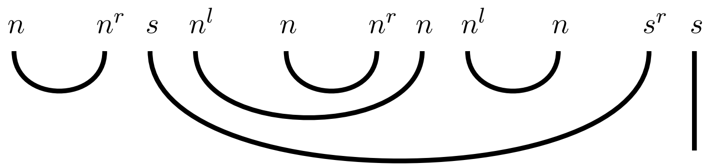
(3) English sentence: "Patrick fought Conor under the large bridge today." In Irish: "Throid Patrick Conor faoin droichead mór inniu".

Patrick fought Conor under the large bridge today

has the type assignment

$$n \quad n^r s n^l \quad n \quad n^r n n^l \quad n \quad s^r s$$

which has a reduction diagram:



In Irish, however,

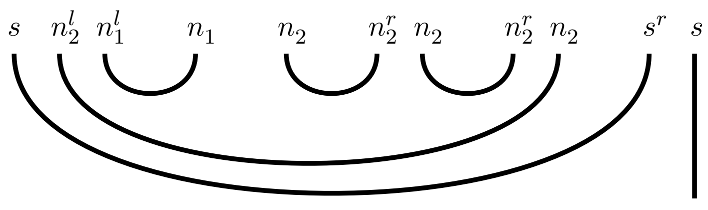
Throid Patrick Conor faoin droichead mór inniu

Fought Patrick Conor under the bridge big today

has the type assignment

$$sn_2^l n_1^l \quad n_1 \quad n_2 \quad n_2^r n_2 \quad n_2^r n_2 \quad s^r s$$

and the reduction diagram



$\Diamond$

[3] "pláta ispíní" = "plate (of) sausages" as a whole prepositional phrase is given the type $n_2^r n_2$.

Finally, Sadrzadeh et al. [27] consider subject relative pronouns (such as *who(m), which*) and object relative pronouns (such as *that*). They assign the pregroup types as follows:

$$n^r n s^l n \text{ (subject relative pronoun)} \qquad n^r n n^{ll} s^l \text{ (object relative pronoun)}$$

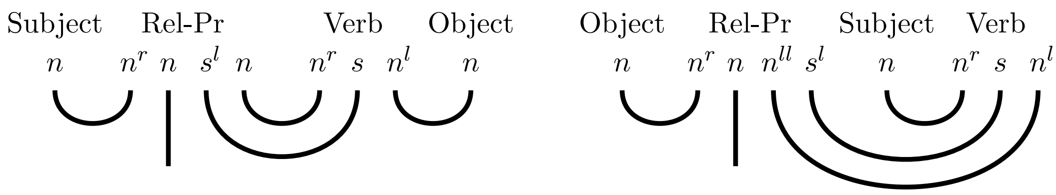


FIGURE 1. Subject relative pronoun. Object relative pronoun.

However, in Irish these particular words (*who(m), which, that*) are simply represented by one word: *a*. Moreover, the grammatical structure of a sentence containing these relative pronouns is the same regardless of whether the relative pronouns are object or subject modifying.

Example 2.3.

men **who** shear sheep
fir **a** chaitheann caorach
men who shear sheep

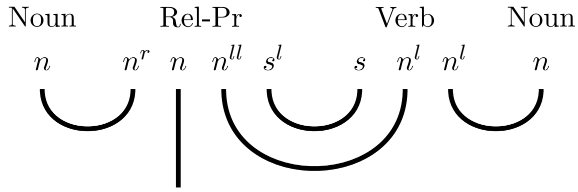
the pig **that** Celia ate
an muc **a** d'ith Celia
the pig that ate Celia

the day **which** was cold
an lá **a** bhí fuar
the day which was cold

◊

So for Irish we can define:

(5) **Relative Pronouns.** Let $n^r n n^{ll} s^l$ be the pregroup type of *a*, the Irish relative pronoun *who(m), which,* and *that*. This results in the following reduction:



This concludes the work required to use a pregroup grammar structure in Irish.

3. A Vector Space based Model of Meaning

The goal of this section is to create a vector space model of meaning from *Corpus A.1*, located in the Appendix. The section after this will create another vector space model of meaning, this time in Irish, from the translation of *Corpus A.1.* The underlying principal is that, once we have the meaning of a sentence in an abstract vector space (S in *Example 2.1*), it does not matter what the language of the sentence is, as it can be compared via an inner product on S. An application of this idea is to measure the accuracy of translation tools such as *Google Translate*, and also to potentially train software (off large corpora) to accept input commands in any language. One could conceive of an extension of this idea to speech recognition, where a speaker of some language utters a sentence, the meaning of which is then calculated as a vector of S, and the command whose meaning is closest to this sentence (the command whose normalised inner product with the meaning of the sentence is closest to 1) being executed. These ideas are beyond the scope of this essay, but our goal is to lay the groundwork here.

The corpus of text chosen by the author is a modified copy of the plot of *Star Wars: Episode III - Revenge of the Sith* obtained from Wikipedia. The full corpus of text is presented in *Appendix A*. We shall closely follow the exposition presented by Grefenstette and Sadrzadeh [13, 15] throughout.

As we are primarily interested in the vector space N of nouns, we shall begin there. We define the basis to consist of the five most commonly occurring words against which we shall measure all other nouns in the corpus:

$$\text{Basis of } N = \{\text{Anakin, Palpatine, Jedi, Obi-Wan, arg-evil}\},$$

where 'arg-evil' denotes the argument of the adjective 'evil' (cf. [14, §3]). The coordinates of a noun K follow from counting the number of times each basis word has appeared in an m word window around K; in particular, K is given a coordinate of k for 'arg-evil' if K has appeared within m words of a noun described as 'evil' in the same sentence, k times in the corpus. For this essay, set $m = 3$. In this basis

$$\text{Anakin} = [1, 0, 0, 0, 0], \tag{1}$$
$$\text{Palpatine} = [0, 1, 0, 0, 0], \tag{2}$$
$$\text{Obi-Wan} = [0, 0, 0, 1, 0], \tag{3}$$
$$\text{Padmé} = [4, 0, 0, 1, 1], \tag{4}$$
$$\text{Yoda} = [0, 1, 1, 3, 1], \tag{5}$$
$$\text{Emperor} = [1, 5, 0, 0, 1], \tag{6}$$
$$\text{mastermind} = [2, 2, 0, 0, 1], \tag{7}$$
$$\text{Mace Windu} = [0, 1, 1, 0, 0], \tag{8}$$
$$\text{Sith Lord} = [1, 1, 0, 0, 1], \tag{9}$$
$$\text{General Grievous} = [0, 1, 3, 1, 0], \tag{10}$$
$$\text{dark side of the Force} = [4, 2, 1, 1, 1], \tag{11}$$

where we treat (11), 'dark side of the Force', as one noun. It has appeared within 3 words of 'Anakin' 4 times, 'Palpatine' 2 times, 'Jedi' once, 'Obi-Wan' once, and the argument of 'evil' once.

As described by Greffenstette and Sadrzadeh [15, Fig. 2] there exists an exact procedure for learning the weights for matrices of words P with relational types π of m

adjoint types. For a given verb V, its weight is

$$C_{ijk} = \begin{cases} \sum_l \sum_{v \in verbs(\mathcal{C}_l)} \delta(v, V) \langle \overrightarrow{subj(v)} | \overrightarrow{n}_i \rangle \langle \overrightarrow{obj(v)} | \overrightarrow{n}_k \rangle \overrightarrow{s}_j & \text{if } \overrightarrow{s}_j = (\overrightarrow{n}_i, \overrightarrow{n}_k), \\ 0 & \text{o.w.} \end{cases} \tag{12}$$

where $\mathcal{C}_l$ is the set of grammatical relations for a sentence s_l in the corpus, $\delta(v, V) = 1$ if $v = V$ and 0 otherwise. As mentioned by Greffenstette et al. [13, 14] if we assume $S = N \otimes N$ (so the basis of S is of the form $(\overrightarrow{n}_i, \overrightarrow{n}_j)$) then the meaning vector of a transitive sentence:

$$\overrightarrow{\text{subject verb object}}$$

is determined by the matrix of the verb, and (12) becomes

$$C_{ik} = \sum_l \sum_{v \in verbs(\mathcal{C}_l)} \delta(v, V) \langle \overrightarrow{subj(v)} | \overrightarrow{n}_i \rangle \langle \overrightarrow{obj(v)} | \overrightarrow{n}_k \rangle. \tag{13}$$

Thus a verb is described by a two dimensional matrix. Using *Corpus A.1*,

$$C^{\text{turn}} = \begin{bmatrix} 10 & 5 & 3 & 2 & 3 \\ 2 & 0 & 0 & 0 & 0 \\ 4 & 2 & 1 & 1 & 1 \\ 0 & 1 & 1 & 3 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

For example, (abbreviating "dark side of the Force" as "DSOF")

$$\begin{aligned} C^{\text{turn}}_{11} &= \sum_l \sum_{v \in verbs(\mathcal{C}_l)} \delta(v, V) \langle \overrightarrow{subj(v)} | \overrightarrow{n}_1 \rangle \langle \overrightarrow{obj(v)} | \overrightarrow{n}_1 \rangle \\ &= \langle \overrightarrow{mastermind} | \overrightarrow{n}_1 \rangle \langle \overrightarrow{Anakin} | \overrightarrow{n}_1 \rangle + \langle \overrightarrow{Anakin} | \overrightarrow{n}_1 \rangle \langle \overrightarrow{Jedi} | \overrightarrow{n}_1 \rangle + \langle \overrightarrow{Jedi} | \overrightarrow{n}_1 \rangle \langle \overrightarrow{DSOF} | \overrightarrow{n}_1 \rangle \\ &+ \langle \overrightarrow{Anakin} | \overrightarrow{n}_1 \rangle \langle \overrightarrow{Palpatine} | \overrightarrow{n}_1 \rangle + 2 \langle \overrightarrow{Anakin} | \overrightarrow{n}_1 \rangle \langle \overrightarrow{DSOF} | \overrightarrow{n}_1 \rangle + \langle \overrightarrow{Anakin} | \overrightarrow{n}_1 \rangle \langle \overrightarrow{evil} | \overrightarrow{n}_1 \rangle \\ &+ \langle \overrightarrow{Obi\text{-}Wan} | \overrightarrow{n}_1 \rangle \langle \overrightarrow{Yoda} | \overrightarrow{n}_1 \rangle \\ &= 2 + 2 \cdot 4 = 10. \end{aligned}$$

We will also require the matrix C^{is} for computations later on. This is again given by equation (13) (where we only use sentences from *Corpus A.1* which have a transitive use of "is", e.g. "Anakin is a powerful Jedi" as opposed to "he is too powerful").

$$C^{\text{is}} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 \\ 4 & 2 & 1 & 1 & 3 \\ 0 & 0 & 1 & 3 & 0 \\ 1 & 1 & 3 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

Of course, an adjective A can be computed in the same fashion:

$$C_{ij} = \begin{cases} \sum_l \sum_{a \in adjs(\mathcal{C}_l)} \delta(a, A) \langle \overrightarrow{arg\text{-}of(a)} | \overrightarrow{n}_i \rangle & \text{if } \overrightarrow{n}_i = \overrightarrow{n}_j \\ 0 & \text{o.w.} \end{cases} \tag{14}$$

This is usually represented as a vector corresponding to the diagonal elements of C; e.g. $C^{powerful} = [1, 3, 1, 3, 1]$ as, for example

$$\begin{aligned} C^{powerful}_{44} &= \sum_l \sum_{a \in adjs(\mathcal{C}_l)} \delta(a, powerful) \langle \overrightarrow{arg\text{-}of(a)} | \overrightarrow{n}_4 \rangle \\ &= 2 \langle \overrightarrow{Palpatine} | \overrightarrow{n}_4 \rangle + \langle \overrightarrow{Anakin} | \overrightarrow{n}_4 \rangle + \langle \overrightarrow{Yoda} | \overrightarrow{n}_4 \rangle \\ &= 3, \end{aligned}$$

or similarly $C^{brave} = [5,1,1,4,1]$ as, for example,

$$\begin{aligned} C^{brave}_{11} &= \sum_l \sum_{a \in adjs(\mathcal{C}_l)} \delta(a, brave) \langle \overrightarrow{arg\text{-}of(a)} | \overrightarrow{n}_1 \rangle \\ &= \langle \overrightarrow{Anakin} | \overrightarrow{n}_1 \rangle + 3 \langle \overrightarrow{Obi\text{-}Wan} | \overrightarrow{n}_1 \rangle + \langle \overrightarrow{Padm\acute{e}} | \overrightarrow{n}_1 \rangle + \langle \overrightarrow{Mace\ Windu} | \overrightarrow{n}_1 \rangle \\ &= 1 + 4 = 5. \end{aligned}$$

3.1. **Warm-up: Representing a sentence as a vector.** Now consider the sentence at the start of *Corpus A.1*:

Palpatine is a mastermind who turns Anakin to the dark side of the Force.

Let use calculate a meaning vector for this sentence. To do this, we first must calculate the corresponding matrix for the prepositional phrase "to the dark side of the Force". This is given by the two dimensional matrix

$$C^{\text{to DSOF}}_{ij} = \begin{cases} \sum_l \sum_{p \in prep(\mathcal{C}_l)} \delta(p, \text{to DSOF}) \langle \overrightarrow{arg\text{-}of(p)} | \overrightarrow{n}_i \rangle & \text{when } \overrightarrow{n}_i = \overrightarrow{n}_j, \\ 0 & \text{o.w.} \end{cases} \tag{15}$$

so

$$C^{\text{to DSOF}} = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

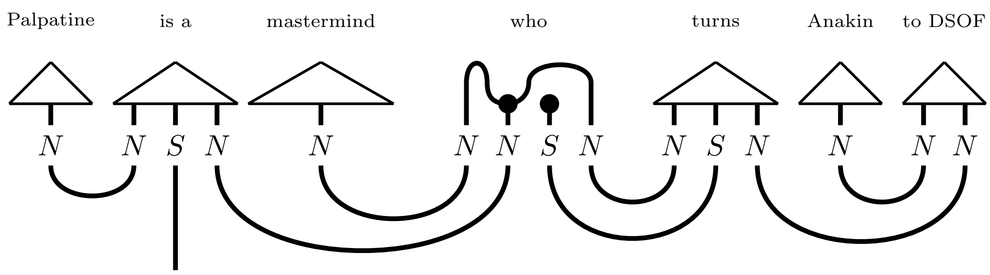
The sentence

Palpatine is a mastermind who turns Anakin to the dark side of the Force

has the type assignment

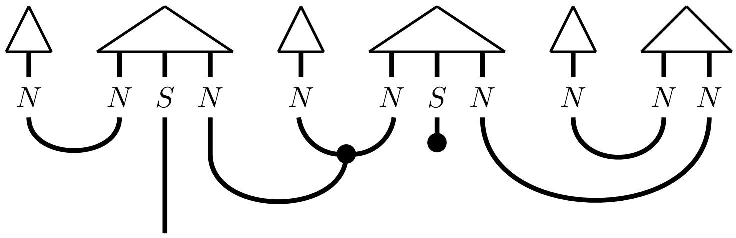
$$n \quad n^r s n^l \quad n \quad n^r n s^l n \quad n^r s n^l \quad n \quad n^r n,$$

using the convention from [13] that the prepositional phrase "to the dark side of the Force" as a whole has the assignment $n^r n$. The reduction diagram for this is[4]:



which, when simplified using the string diagram rules of [27], becomes

[4]Using Frobenius algebras as outlined by Sadradeh et al. [27], and abbreviating "the dark side of the Force" as "DSOF".



The corresponding map for this reduction diagram is

$$f = (1_S \otimes \epsilon_N \otimes \epsilon_N) \circ (\epsilon_N \otimes 1_S \otimes 1_N \otimes \mu_N \otimes i_S \otimes 1_N \otimes \epsilon_N \otimes 1_N), \tag{16}$$

so

$$\begin{aligned}
&\overrightarrow{\text{Palpatine is a mastermind who turns Anakin to the dark side of the Force}} \\
&= f(\overrightarrow{\text{Palpatine}} \otimes \overrightarrow{\text{is}} \otimes \overrightarrow{\text{mastermind}} \otimes \overrightarrow{\text{turns}} \otimes \overrightarrow{\text{Anakin}} \otimes \overrightarrow{\text{to DSOF}}) \\
&= f\left(\left(\sum_k c_k^{\text{Palp}} \overrightarrow{n}_k \right) \otimes \left(\sum_{lpq} c_{lpq}^{\text{is}} \overrightarrow{n}_l \otimes \overrightarrow{s}_p \otimes \overrightarrow{n}_q \right) \otimes \left(\sum_r c_r^{\text{mm}} \overrightarrow{n}_r \right) \right. \\
&\left. \otimes \left(\sum_{stu} c_{stu}^{\text{turns}} \overrightarrow{n}_s \otimes \overrightarrow{s}_t \otimes \overrightarrow{n}_u \right) \otimes \left(\sum_v c_v^{\text{Anakin}} \overrightarrow{n}_v \right) \otimes \left(\sum_{wx} c_{wx}^{\text{to DSOF}} \overrightarrow{n}_w \otimes \overrightarrow{n}_x \right) \right) \\
&= \sum_{k,p,r,v,x} c_k^{\text{Palp}} c_{kpr}^{\text{is}} c_r^{\text{mm}} c_{rx}^{\text{turns}} c_v^{\text{Anakin}} c_{vx}^{\text{to DSOF}} \overrightarrow{s}_p \\
&= \sum_p (60 c_{2p1}^{\text{is}} + 12 c_{2p2}^{\text{is}} + 3 c_{2p5}^{\text{is}}) \overrightarrow{s}_p. \qquad (\clubsuit)
\end{aligned}$$

This does not have much meaning, as S is an arbitrary vector space. If we set $S = N \otimes N$, then $\overrightarrow{s}_p = (\overrightarrow{n}_i, \overrightarrow{n}_j)$, and the verb matrix C^{is} becomes

$$C_{ijk}^{\text{is}} = \begin{cases} \sum_l \sum_{v \in verbs(\mathcal{C}_l)} \delta(v, V) \langle \overrightarrow{subj(v)} | \overrightarrow{n}_i \rangle \langle \overrightarrow{obj(v)} | \overrightarrow{n}_k \rangle \overrightarrow{s}_j & = C_{ik}^{\text{is}} \quad \text{if } \overrightarrow{s}_j = \overrightarrow{n}_i \otimes \overrightarrow{n}_j \\ & 0 \quad \text{o.w.,} \end{cases}$$

by equation (13). Thus

$$(\clubsuit) = 240 \overrightarrow{n}_2 \otimes \overrightarrow{n}_1 + 24 \overrightarrow{n}_2 \otimes \overrightarrow{n}_2 + 9 \overrightarrow{n}_2 \otimes \overrightarrow{n}_5.$$

One could read into this by arguing the sentence "Palpatine is a mastermind who turns Anakin to the dark side of the Force" is a combination of (Palpatine, Anakin), (Palpatine, Palpatine), and (Palpatine, evil) but really this sum of tensor products only becomes meaningful when we are comparing sentences via an inner product on S, as we do in the next section.

3.2. **Sentence Comparison.** Consider the sentences

(1) Yoda is a powerful Jedi.
(2) Obi-Wan is a brave Jedi.
(3) Palpatine is a brave Jedi.

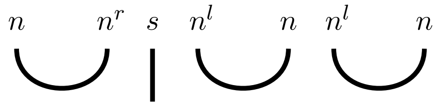
To compute the meaning of these we need the reduction diagram for a sentence of the form

noun is adjective noun

which has the type assignment

$$n \quad n^r s n^l \quad n n^l \quad n.$$

The reduction diagram is:



which corresponds to a map $f = \epsilon_N \otimes 1_S \otimes \epsilon_N \otimes \epsilon_N$. Therefore

$$\begin{aligned}
&\overrightarrow{\text{X is a Y Z}} = f(\overrightarrow{\text{X}} \otimes \overrightarrow{\text{is}} \otimes \overrightarrow{\text{Y}} \otimes \overrightarrow{\text{Z}}) \\
&= (\epsilon_N \otimes 1_S \otimes \epsilon_N \otimes \epsilon_N)\left(\left(\sum_i c_i^{\text{X}} \overrightarrow{n}_i\right) \otimes \left(\sum_{jkl} c_{jkl}^{\text{is}} \overrightarrow{n}_j \otimes \overrightarrow{s}_k \otimes \overrightarrow{n}_l\right) \otimes \left(\sum_{pq} c^{\text{Y}} \overrightarrow{n}_p \otimes \overrightarrow{n}_q\right) \otimes \left(\sum_r c_r^{\text{Z}} \overrightarrow{n}_r\right)\right) \\
&= \sum_{jkl,p} c_j^{\text{X}} c_{jkl}^{\text{is}} c_{lp}^{\text{Y}} c_p^{\text{Z}} \overrightarrow{s}_k.
\end{aligned}$$

Compare the sentences "Yoda is a powerful Jedi" and "Obi-Wan is a brave Jedi".

$$\begin{aligned}
&\langle \text{Yoda is a powerful Jedi} \mid \text{Obi-Wan is a brave Jedi} \rangle \\
&= \left\langle \sum_{jkl,p} c_j^{\text{Yoda}} c_{jkl}^{\text{is}} c_{lp}^{\text{powerful}} c_p^{\text{Jedi}} \overrightarrow{s}_k \,\middle|\, \sum_{jkl,p} c_j^{\text{Obi-Wan}} c_{jkl}^{\text{is}} c_{lp}^{\text{brave}} c_p^{\text{Jedi}} \overrightarrow{s}_k \right\rangle \\
&= 27.
\end{aligned}$$

Normalise this by dividing by the square root of the product of the length of the two sentences:

$$\begin{aligned}
\langle \text{Yoda is a powerful Jedi} \mid \text{Yoda is a powerful Jedi} \rangle &= 84, \\
\langle \text{Obi-Wan is a brave Jedi} \mid \text{Obi-Wan is a brave Jedi} \rangle &= 9,
\end{aligned}$$

We get that the sentences "Yoda is a powerful Jedi" and "Obi-Wan is a brave Jedi" have a similarity score of $\frac{29}{\sqrt{84 \cdot 9}} = 0.9812$; very high. On the other hand, the inner product of the sentences

$$\begin{aligned}
&\langle \text{Yoda is a powerful Jedi} \mid \text{Palpatine is a brave Jedi} \rangle \\
&= \sum_{jl,p} c_j^{\text{Yoda}} c_j^{\text{Palpatine}} c_{jl}^{\text{is}^2} c_{lp}^{\text{powerful}} c_{lp}^{\text{brave}} c_p^{\text{Jedi}^2} \\
&= 1,
\end{aligned}$$

and their lengths are

$$\begin{aligned}
\langle \text{Yoda is a powerful Jedi} \mid \text{Yoda is a powerful Jedi} \rangle &= 84, \\
\langle \text{Palpatine is a brave Jedi} \mid \text{Palpatine is a brave Jedi} \rangle &= 1,
\end{aligned}$$

hence their normalised similarity score is $\frac{1}{\sqrt{1 \cdot 84}} = 0.1091$; quite low.

The reason for these scores is that, in the corpus, Obi-Wan and Yoda are referred to as brave and powerful Jedi, whereas Palpatine is never referred to as a Jedi, only as a powerful, evil Sith Lord or mastermind.

If we consider the sentence

The Emperor is a mastermind who turns Anakin to the dark side of the Force

which, by *Section 3.1*, has a meaning vector

$$\overrightarrow{\text{The Emperor is a mastermind who turns Anakin to the dark side of the Force}}$$

$$= f(\overrightarrow{\text{Emperor}} \otimes \overrightarrow{\text{is}} \otimes \overrightarrow{\text{mastermind}} \otimes \overrightarrow{\text{turn}} \otimes \overrightarrow{\text{Anakin}} \otimes \overrightarrow{\text{to DSOF}}) \qquad (f \text{ given by (16)})$$
$$= \sum_{k,p,r,v,x} c_k^{\text{Emp}} c_{kpr}^{\text{is}} c_r^{\text{mm}} c_{rx}^{\text{turn}} c_v^{\text{Anakin}} c_{vx}^{\text{to DSOF}} \overrightarrow{s}_p$$
$$= 60\overrightarrow{n}_1 \otimes \overrightarrow{n}_1 + 3\overrightarrow{n}_1 \otimes \overrightarrow{n}_5 + 1200\overrightarrow{n}_2 \otimes \overrightarrow{n}_1 + 2\overrightarrow{n}_2 \otimes \overrightarrow{n}_2 + 45\overrightarrow{n}_2 \otimes \overrightarrow{n}_5 + 3\overrightarrow{n}_5 \otimes \overrightarrow{n}_5,$$

when we compare this sentence to the one at the beginning of *Corpus A.1*,

$$\langle \overrightarrow{\text{The Emperor is a mastermind who turns Anakin to the dark side of the Force}} \mid \overrightarrow{\text{Palpatine is a mastermind who turns Anakin to the dark side of the Force}} \rangle = 288453.$$

The length of the former is 1445647, and the length of the latter is 58257. Therefore their similarity score is $\frac{288453}{\sqrt{1445647 \cdot 58257}} = 0.9939$; very high. Of course, as Palpatine is the Emperor, it should be very high!

A similarly quick calculation of the inner product of the sentences "Padmé is a mastermind who turns Anakin to the dark side of the Force" and "Palpatine is a mastermind who turns Anakin to the dark side of the Force" gives a similarity score of 0; as expected these sentences are not similar at all, as "Padmé" is very different to "Palpatine".

The vector space model of meaning has managed to extract these key themes from the corpus. Now our goal is to extract the same key ideas from an Irish corpus.

4. Bilingual Sentence Comparison via the Vector Space Model of Meaning

We shall now compare sentences between corpora in different languages. Our Irish vector space model of meaning shall be created from *Corpus B.1*, using the methods detailed in the previous section.

The calculations in *Section 3* required $S = N \otimes N$, however this becomes a problem moving between languages; the noun space in the Irish model of the meaning, denoted N', is a different space to the noun space N of the English vector space model. However, if we assume the bases of N and N' are the same, then the basis of S will still be $\{(\overrightarrow{n}_i, \overrightarrow{n}_j)\}$ meaning the inner product on S can still be computed as it was in *Section 3* and [14]. To that end, let the basis of N' be

$$\{\text{Anakin}, \text{Palpatine}, \text{Jedi}, \text{Obi-Wan}, \text{arg-olc}\},$$

where "arg-olc" corresponds to the argument for the adjective olc - in English, 'evil'. This is also the collection of the five most commonly occurring nouns in *Corpus B.1* exactly (which might not really be a surprise as *Corpus B.1* is a translation of *Corpus A.1*, and nouns in English typically have one translation to Irish).

Take for example the sentence "Yoda is a powerful Jedi". In Irish, this is "Is Jedi cumhachtach é Yoda". Translated literally, it becomes "Is Jedi powerful Yoda" - amusingly, closer to Yoda's speech pattern than to English. Using DisCoCat models, we get promising results:

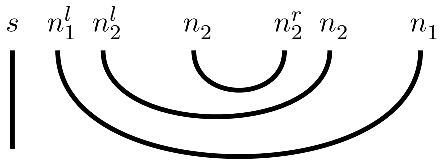
The sentence

Is Jedi cumhachtach é Yoda

has the type assignment

$$s n_1^l n_2^l \quad n_2 \quad n_2^r n_2 \quad n_1$$

and the reduction diagram



corresponding to a map

$$f = (1_S \otimes \epsilon_N) \circ (1_S \otimes 1_n \otimes \epsilon_N \otimes 1_N) \circ (1_S \otimes 1_N \otimes 1_N \otimes \epsilon_N \otimes 1_N \otimes 1_N).$$

Therefore the sentence "Is Jedi cumhachtach é Yoda" is assigned the following meaning vector:

$$\begin{aligned}
&\overrightarrow{\text{Is Jedi cumhachtach é Yoda}}\\
&= f(\overrightarrow{\text{Is}} \otimes \overrightarrow{\text{Jedi}} \otimes \overrightarrow{\text{cumhachtach}} \otimes \overrightarrow{\text{Yoda}})\\
&= f\left(\left(\sum_{jkl} c^{\text{Is}}_{ijk} \overrightarrow{s}_i \otimes \overrightarrow{n}_j \otimes \overrightarrow{n}_k\right) \otimes \left(\sum_l c^{\text{Jedi}}_l \overrightarrow{n}_l\right) \otimes \left(\sum_{pq} c^{\text{cumh}}_{pq} \overrightarrow{n}_p \otimes \overrightarrow{n}_q\right) \otimes \left(\sum_r c^{\text{Yoda}}_r \overrightarrow{n}_r\right)\right)\\
&= \sum_{ijk,p} c^{\text{Is}}_{ijk} c^{\text{Jedi}}_p c^{\text{cumh}}_{pk} c^{\text{Yoda}}_j \overrightarrow{s}_i.
\end{aligned}$$

In order to evaluate this sentence, we need values for $c^{\text{Is}}_{ijk}, c^{\text{Jedi}}_p, c^{\text{cumh}}_{pk}$, and c^{Yoda}_j. These are calculate in the same way as in *Section 3*, using equations (13), (14), and (15) as well as the 3 word window to assign coordinates to nouns. In particular the matrix C^{Is} is calculated as follows: the copula "is" is translated to have the same meaning as the verb "to be" in English, which in Irish corresponds to the verb "bí", which in *Corpus B.1* is conjugated as "tá". The result? $C^{\text{Is}} = C^{\text{tá}}$ is calculated by including sentences with use of either "Tá. . . " or "Is. . . ".

$$\begin{aligned}
\text{Jedi} &= [0, 0, 1, 0, 0],\\
\text{Yoda} &= [0, 1, 2, 3, 0],\\
\text{taobh dorcha na Fórsa}^5 &= [4, 2, 0, 0, 1],\\
C^{\text{cumh}} &= [1, 3, 2, 3, 0],\\
C^{\text{Is}} &= \begin{bmatrix} 4 & 0 & 1 & 0 & 1\\ 4 & 6 & 1 & 0 & 2\\ 1 & 0 & 3 & 0 & 2\\ 0 & 0 & 3 & 0 & 0\\ 1 & 0 & 0 & 0 & 1 \end{bmatrix},
\end{aligned}$$

for example

$$\begin{aligned}
C^{\text{Is}}_{21} &= \sum_l \sum_{v \in verbs(\mathcal{C}_l)} \delta(v, V) \langle \overrightarrow{subj(v)} | \overrightarrow{n}_2 \rangle \langle \overrightarrow{obj(v)} | \overrightarrow{n}_1 \rangle\\
&= \langle \overrightarrow{\text{taobh dorcha na Fórsa}} | \overrightarrow{n}_2 \rangle \langle \overrightarrow{\text{Anakin}} | \overrightarrow{n}_1 \rangle + 2 \langle \overrightarrow{\text{Palpatine}} | \overrightarrow{n}_2 \rangle \langle \overrightarrow{\text{máistirmind}} | \overrightarrow{n}_1 \rangle\\
&= 2 + 2 = 4.
\end{aligned}$$

[5] "dark side of the Force".

It is quite welcome that the vectors $\overrightarrow{\text{Yoda}}$ and $\overrightarrow{\text{Yoda}}$ are distinct; the grammar of Irish changes the word order of sentences from English, hence (for example) Yoda occurs more frequently with Palpatine in *Corpus B.1* than *Corpus A.1*.

Note that C^{Is} is different to the English C^{is}, as in Irish the verb "to be" is sometimes used in conjunction with another verb, which becomes the main transitive verb of the sentence. Thus, there are fewer occurrences of "tá" or "is" in *Corpus B.1* than "is" in *Corpus A.1*.

The result of our two assumptions (that $S = N \otimes N$ and the basis of N' is the exact translation of the basis of N) is we can meaningfully compare the following sentences:

$$
\begin{aligned}
&\langle \text{Yoda is a powerful Jedi} \mid \text{Is Jedi cumhachtach é Yoda} \rangle \\
&= \left\langle \sum_{jkl,p} c_j^{\text{Yoda}} c_{jkl}^{\text{is}} c_{lp}^{\text{powerful}} c_p^{\text{Jedi}} \vec{s}_k \,\middle|\, \sum_{jkl,p} c_{kjl}^{\text{Is}} c_p^{\text{Jedi}} c_{pl}^{\text{cumh}} c_j^{\text{Yoda}} \vec{s}_k \right\rangle \\
&= \sum_{jl,p} c_{jl}^{\text{is}} c_{jl}^{\text{Is}} c_j^{\text{Yoda}} c_j^{\text{Yoda}} c_p^{\text{Jedi}} c_p^{\text{Jedi}} c_{lp}^{\text{powerful}} c_{pl}^{\text{cumh}} \\
&= 2 \sum_j c_{j3}^{\text{is}} c_{j3}^{\text{Is}} c_j^{\text{Yoda}} c_j^{\text{Yoda}} \\
&= 176.
\end{aligned}
$$

To normalise this we calculate

$$
\langle \text{Is Jedi cumhachtach é Yoda} \mid \text{Is Jedi cumhachtach é Yoda} \rangle = 472,
$$
$$
\langle \text{Yoda is a powerful Jedi} \mid \text{Yoda is a powerful Jedi} \rangle = 84,
$$

meaning the similarity score between the sentence "Yoda is a powerful Jedi" and its Irish translation "Is Jedi cumhachtach é Yoda" is $\frac{176}{\sqrt{84 \cdot 472}} = 0.884$; high.

On the other hand, if we try to compare sentences that are not translates of one another, say "Is Jedi cróga é Palpatine" (in English, "Palpatine is a brave Jedi"), we receive low scores:

$$
C^{\text{cróga}} = [4, 1, 1, 4, 0] \qquad \text{by equation (14)}^6.
$$

$$
\begin{aligned}
&\langle \text{Yoda is a powerful Jedi} \mid \text{Is Jedi cróga é Palpatine} \rangle \\
&= \left\langle \sum_{jkl,p} c_j^{\text{Yoda}} c_{jkl}^{\text{is}} c_{lp}^{\text{powerful}} c_p^{\text{Jedi}} \vec{s}_k \,\middle|\, \sum_{jkl,p} c_{kjl}^{\text{Is}} c_p^{\text{Jedi}} c_{pl}^{\text{cróga}} c_j^{\text{Palpatine}} \vec{s}_k \right\rangle \\
&= 1.
\end{aligned}
$$

To normalise this we calculate

$$
\langle \text{Is Jedi cróga é Palpatine} \mid \text{Is Jedi cróga é Palpatine} \rangle = 1,
$$
$$
\langle \text{Yoda is a powerful Jedi} \mid \text{Yoda is a powerful Jedi} \rangle = 84,
$$

meaning the similarity score between the sentence "Yoda is a powerful Jedi" and "Is Jedi cróga é Palpatine" is $\frac{1}{\sqrt{84 \cdot 1}} = 0.1091$; quite low.

There is one problem: the sentences "Yoda is a powerful Jedi" and "Is Tiarna Sith cumhachtach é Yoda" (in English, "Yoda is a powerful Sith Lord") have a high similarity score.

---

[6] Technically when calculating $C_{ij}^{\text{cróga}}$ we are also counting the various different translations of "brave" occuring in *Corpus B.1*, such as "go crua" or "go láidir".

Example 4.1. From *Corpus B.1*,

$$\text{Tiarna Sith}^7 = [1,0,1,0,1],$$

$$\begin{aligned}
&\langle \text{Yoda is a powerful Jedi} \mid \text{Is Tiarna Sith cumhachtach é Yoda} \rangle \\
&= \left\langle \sum_{jkl,p} c_j^{\text{Yoda}} c_{jkl}^{\text{is}} c_{lp}^{\text{powerful}} c_p^{\text{Jedi}} \overrightarrow{s}_k \middle| \sum_{jkl,p} c_{kjl}^{\text{Is}} c_p^{\text{Sith}} c_{pl}^{\text{cumh}} c_j^{\text{Yoda}} \overrightarrow{s}_k \right\rangle \\
&= 176,
\end{aligned}$$

when normalised by

$$\langle \text{Is Tiarna Sith cumhachtach é Yoda} \mid \text{Is Tiarna Sith cumhachtach é Yoda} \rangle = 472,$$
$$\langle \text{Yoda is a powerful Jedi} \mid \text{Yoda is a powerful Jedi} \rangle = 84,$$

becomes $\frac{176}{\sqrt{84 \cdot 472}} = 0.884$; high. ♢

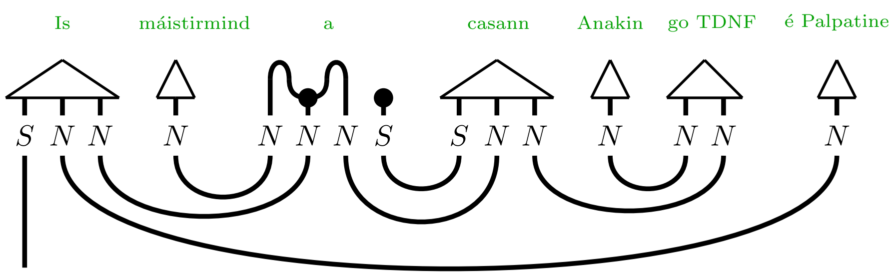
This is because in this model "Tiarna Sith" and "Jedi" are quite similar as vectors, the former being $[1,0,1,0,1]$ and the latter $[0,0,1,0,0]$. In English "Sith Lord" was given the vector $[1,1,0,0,1]$ hence is not as easily confused with "Jedi". This means our Irish model has a slightly different idea of what a "Tiarna Sith" is, compared to the English model; in Irish "Tiarna Sith" is closer to "Jedi" than "Sith Lord" is to "Jedi".

Finally, we preform one last grand example.

4.1. **A Complicated Translation.** To conclude this section we shall compare the similarity of meaning between "Palpatine is a mastermind who turns Anakin to the dark side of the Force" and its Irish equivalent, "Is máistirmind a casann Anakin go taobh dorcha na Fórsa é Palpatine". The Irish sentence is assigned the following type:

Is	máistirmind	a	casann	Anakin	go taobh dorcha na Fórsa	é Palpatine
$sn_1^l n_2^l$	n_2	$n_2^r n_2 n_2^{ll} s^l$	$sn_2^l n_1^l$	n_1	$n_1^r n_1$	n_1
Is a	mastermind	who	turns	Anakin	to side dark of the Force	Palpatine

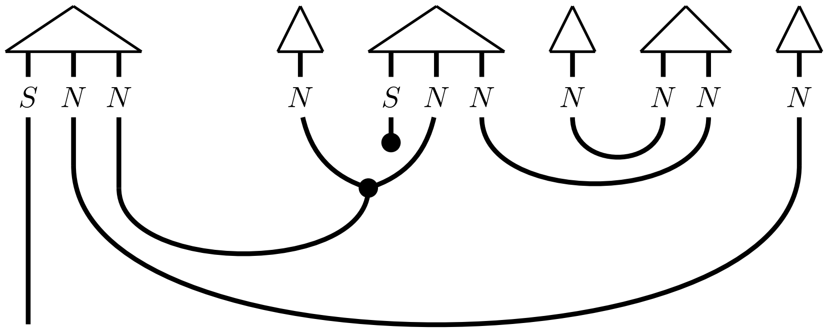
Abbreviating "taobh dorcha na Fórsa" as "TDNF", the reduction diagram is[8]:



which when simplified becomes:

[7]"Sith Lord".

[8]Taking cues from the English "who" [27] regarding the depiction of "a" in the diagram.



and corresponds to a map,

$$f = (1_S \otimes \epsilon_N) \circ (1_S \otimes 1_N \otimes \epsilon_N \otimes 1_N) \circ (1_S \otimes 1_N \otimes 1_N \otimes \mu_N \otimes \epsilon_N \otimes 1_N) \\ \circ (1_S \otimes 1_N \otimes 1_N \otimes 1_N \otimes i_S \otimes 1_N \otimes 1_N \otimes \epsilon_N \otimes 1_N \otimes 1_N).$$

Since

$$\text{máistirmind} = [1, 3, 0, 0, 1],$$

the meaning vector of the sentence is:

$$\begin{aligned} &\overrightarrow{\text{Is máistirmind a casann Anakin go taobh dorcha na Fórsa é Palpatine}} \\ &= f(\overrightarrow{\text{Is}} \otimes \overrightarrow{\text{máistirmind}} \otimes \overrightarrow{\text{casann}} \otimes \overrightarrow{\text{Anakin}} \otimes \overrightarrow{\text{go TDNF}} \otimes \overrightarrow{\text{Palpatine}}) \\ &= \sum_{i,j,k,r,s} c^{\text{is}}_{ijk} c^{\text{mm}}_{k} c^{\text{casann}}_{kr} c^{\text{Anakin}}_{s} c^{\text{go TDNF}}_{sr} c^{\text{Palp}}_{j} \overrightarrow{s}_i \\ &= \sum_{i,k} 3(10 c^{\text{is}}_{i21} + 3 c^{\text{is}}_{i22} + 2 c^{\text{is}}_{i25}) \overrightarrow{s}_i \qquad (\spadesuit) \end{aligned}$$

We don't need to calculate the full matrices C^{casann} and $C^{\text{go TDNF}}$, only the relevant parts, which the author has excluded for brevity[9]. These are calculated as per (13) and (15). So

$$(\spadesuit) = 120 \overrightarrow{n}_2 \otimes \overrightarrow{n}_1 + 54 \overrightarrow{n}_2 \otimes \overrightarrow{n}_2 + 12 \overrightarrow{n}_2 \otimes \overrightarrow{n}_5.$$

Taking the inner product,

$$\begin{aligned} &\langle \overrightarrow{\text{Palpatine is a mastermind who turns Anakin to the dark side of the Force}} \mid \\ &\quad \overrightarrow{\text{Is máistirmind a casann Anakin go taobh dorcha na Fórsa é Palpatine}} \rangle \\ &= 30204. \end{aligned}$$

The length of the former is 57825, and the length of the latter is 17460. Therefore their similarity score is $\frac{30204}{\sqrt{57825 \cdot 17460}} = 0.9506$; very high.

Suppose we thought the translation of "Is máistirmind a casann Anakin go taobh dorcha na Fórsa é Palpatine" was "*The Emperor* is a mastermind who turns Anakin to the dark side of the Force". Using the sentence vector

$$\begin{aligned} &\overrightarrow{\text{The Emperor is a mastermind who turns Anakin to the dark side of the Force}} \\ &= 60 \overrightarrow{n}_1 \otimes \overrightarrow{n}_1 + 3 \overrightarrow{n}_1 \otimes \overrightarrow{n}_5 + 1200 \overrightarrow{n}_2 \otimes \overrightarrow{n}_1 + 2 \overrightarrow{n}_2 \otimes \overrightarrow{n}_2 + 45 \overrightarrow{n}_2 \otimes \overrightarrow{n}_5 + 3 \overrightarrow{n}_5 \otimes \overrightarrow{n}_5, \end{aligned}$$

[9] In particular, $c^{\text{casann}}_{11} = 10, c^{\text{casann}}_{21} = 1, c^{\text{casann}}_{51} = 2$ and $c^{\text{go TDNF}}_{11} = 3$.

from *Section 3.2*, we can calculate:

$$\langle \overrightarrow{\text{The Emperor is a mastermind who turns Anakin to the dark side of the Force}} \mid$$
$$\overrightarrow{\text{Is máistirmind a casann Anakin go taobh dorcha na Fórsa é Palpatine}} \rangle$$
$$= 144648.$$

The length of the former is 1449243, and the length of the latter is 17460. Therefore the similarity score of the sentences is $\frac{144648}{\sqrt{1449243 \cdot 17460}} = 0.9093$; high, but not as high as the actual translation.

Based on this exercise and the calculations of *Section 4*, the author recommends setting a threshold similarity score of 0.8, i.e. 80%: if two sentences (one in English, the other in Irish) are 80% or more similar, they can be deemed as translations of one another relative to the underlying corpora.

Of course, this means that *Example 4.1*, "Yoda is a powerful Jedi" and "Is Tiarna Sith cumhachtach é Yoda" are translations of one another - which is not ideal. In the remainder of the essay we shall work with *conceptual spaces*; instead of nouns being labelled relative to nouns they appear often with, instead nouns are represented by other words that describe them. The hope is this removes instances like the aforementioned problematic translation. However, building on the ideas of Bolt et al. [3] and Gärdenfors [10, 11, 12] much work would need to be done to capture the intricacies between Sith Lords and Jedi Knights. Instead, we will first tackle the problem of automatically creating conceptual spaces for simpler, more distinct nouns such as fruits and planets.

5. Word Classification

According to Dixon and Aikhenvald [8], "three word classes are . . . implicit in the structure of each human language: nouns, verbs and adjectives." It is the goal of this section to specify a treatment of nouns, verbs and adjectives for use in conceptual space creation. Once we have some sort of classification system for each of these, we can proceed with automatically creating a conceptual space from a given corpus. For example, in the case of adjectives we wish to classify words such has 'heavy', 'red' or 'hot', and to each assign a numeric value that transcends language and thus can be compared across (say) Irish and English.

5.1. **Adjectives.**

In their landmark work, Dixon and Aikhenvald [8] give a complete treatment of adjective classes as they arise in various languages across the globe, such as Japanese, Korean, Jarawara, Mam and Russian. In particular, they name seven core types of adjectives that consistently and naturally arise:

(1) **Dimension.** (big, small, short, tall, etc.)
(2) **Age.** (new, old, etc.)
(3) **Value.** (good, bad, curious, necessary, expensive, etc.)
(4) **Colour.** (green, white, orange, etc.)
(5) **Physical Property.** (hard, hot, heavy, wet, soft, etc.)
(6) **Human Propensity.** (kind, happy, sad, greedy, etc.)
(7) **Speed.** (fast, slow, etc.)

As our focus will be representing (non-human) nouns as conceptual spaces, we will not consider item (6). Also, as our focus will be on recreating a human's process of conceptual space construction, the author proposes reframing some of these seven core adjective types from the perspective of our five senses; sight, smell, sound, sensation and savour:

(1) **Dimension.**
(2) **Age.**
(3) **Value.**
(4) **Physical Property.** Further classified as:
   (a) Colour, Intensity (Sight)
   (b) Smell
   (c) Savour (Taste)
   (d) Sound
   (e) Temperature, Density, Mass and Texture (Sensation)
(5) **Speed.**

How do we represent this data numerically? Fortunately most aspects of the five categories lend themselves to a linear interpretation. For example, in **Dimension** we can order adjectives in this class from 'small' to 'large' and represent **Dimension** as an interval $[0, 1]$. This will not be extremely precise - nor, in fact, do we want it to be - by our very nature spaces visualised by humans are fuzzy, and our use of adjectives reflects this. One can equally describe a quadruple patty burger, and the Sun, as 'huge'; maybe 'huge' is 0.9 on the $[0, 1]$ dimension scale. On the other hand, in reality one is far larger than the other. The beauty of DisCoCat models, however, is the context of a sentence feeds the semantic interpretation of the sentence. If we are in the context of food or astronomical bodies, this is a fine figure to assign 'huge', as relative to those two subjects those items are 'huge'. If we find ourselves in a context where both food and astronomical bodies are being discussed, it is true that things might become more jumbled. Consider the sentence:

"It is a fact of biology and physics that a quadruple patty burger is a huge portion of food, and the Sun is a huge astronomical body."

In this case we posit that, should the corpus be longer, natural speakers of the language will use different adjectives when providing a more complete description of quadruple patty burgers and the Sun, so our argument for allowing **Dimension** to be ordered without context is reasonable. Of course, if the *only* description one had ever heard about quadruple patty burgers and the Sun is "it is a fact of biology and physics that a quadruple patty burger is a huge portion of food, and the Sun is a huge astronomical body" it is more than fair to confuse the size of the two!

Similarly we allow **Age** to be represented by $[0, 1]$ (where adjectives such as *young, new, baby* are closer to 0, and *old, mature, antiquated* are closer to 1), and **Value** and **Speed** to be represented by $[0, 1]$ as well[10]. We will take these spaces as given and assume one can preload a list of common adjectives with assigned $[0, 1]$ values, in much the same way it is assumed one can preload a list of colours with assigned $[0, 1]^3$ values in the common RGB colour cube.

For **Physical Properties**,

(a) Colour will be represented numerically by the RGB colour cube, and Intensity by the interval $[0, 1]$.
(b) Smell we shall represent by Henning's Prism; published by Hans Henning in 1915, it classifies odours according to six primary odours: fragrant, ethereal, resinous, spicy, putrid, and burnt (*See figure 2 on the following page*).

    Embed this into $\mathbb{R}^3$ in the usual way: Fragrant $= [1, -1, 0]$, Ethereal $= [-1, -1, 0]$, Spicy $= [1, 1, 0]$, Resinous $= [-1, 1, 0]$, Putrid $= [0, -1, 1]$ and Burned $= [0, 1, 1]$.

[10]In the case of **Value**, we order words from 'low value' - such as *inexpensive, bad, fake* - to 'high value' - such as *necessary, crucial, costly.*

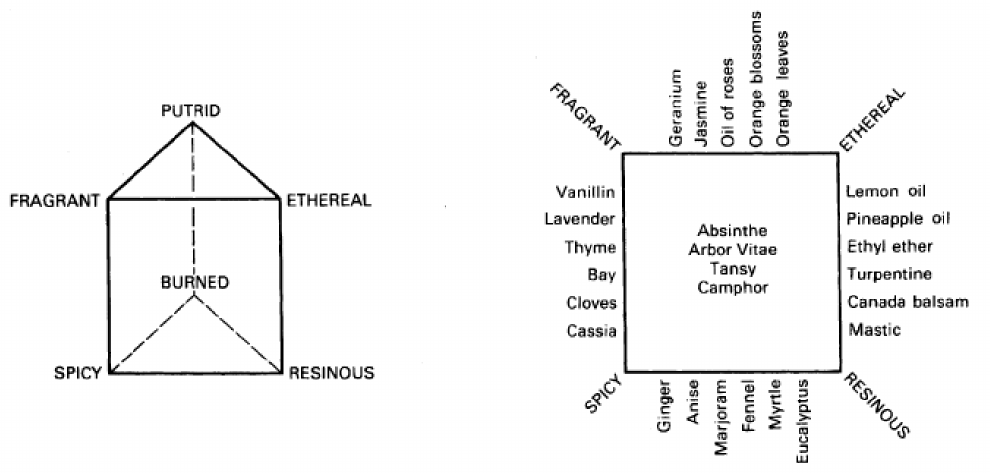


FIGURE 2. Henning's Smell Prism, courtesy of [22].

(c) Savour by Gärdenfors' taste tetrahedron:

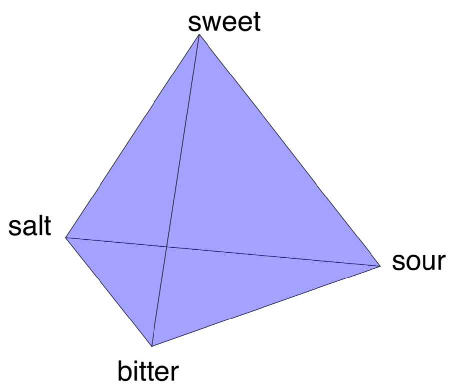


FIGURE 3. Gärdenfors Taste Tetrahedron, courtesy of [3].

Embed this into $\mathbb{R}^3$ in the usual way: Salt $= [1, 0, 0]$, Sour $= [-\frac{1}{2}, -\frac{\sqrt{3}}{2}, 0]$, Bitter $= [-\frac{1}{2}, \frac{\sqrt{3}}{2}, 0]$, and Sweet $= [0, 0, \sqrt{2}]$.

(d) The author has no firm suggestions for representing sound, though the work of Forth et al. [9] describes how a range of musical qualities may be described through conceptual spaces. For a minimal working example in this essay, the author suggests using a square $[0, 1]^2$ where the first dimension represents intensity (from quiet to loud) and the second dimension represents feeling towards the sound (from bad to indifferent/undefined to good). So if a sound was described as "muffled and pleasant" it could be assigned the point $(0.1, 0.9)$, whereas a noise reported as "loud" would be assigned the point $(0.9, 0.5)$.

(e) Sensation we can represent by a hypercube $[0, 1]^4$ with the first dimension temperature (from low to high), the second dimension density (from low - e.g. *gaseous, wispy, fine*, to high - e.g. *solid, dense, hard*, with items like *soft, mushy, liquidy, wet, gloopy, sticky, brittle, crumbly* in between), the third dimension mass (from light to heavy) and the fourth dimension texture (from smooth to rough).

This system cannot capture every type of adjective (in particular, *texture* leaves much to be desired). Also, at present this view is not sophisticated enough to capture 'dry', 'clear', 'sunny' etc., and *density* seems overloaded with information. However, this system is sufficiently complex and complete to allow us to start analysing text in a meaningful way. Going forward, we shall assign numerical values to adjectives based on our intuition and assume a complex set of adjectives has been hard coded into our algorithm a priori. This may seem a little ad hoc, but it is how we learn adjectives in the early years of our life; by repeated exposure and memorisation.

Of course, our mental picture of objects comes not just from adjectives, but also other nouns.

5.2. **Nouns.** The advantage to allowing nouns to classify other nouns is twofold; first, nouns can identify structure that adjectives might have missed. For example, describing apples and cars as "red, smooth and fresh smelling" might be accurate, but paints the wrong conceptual picture. The picture is corrected once we include the sentences "an apple is a fruit" and "a car is a vehicle". Such classifying words as 'fruit' or 'vehicle' are known as *hypernyms*; a word A is a hypernym of a word B if the sentence "B is a (kind of) A" is acceptable to English speakers. The converse, a *hyponym*, is defined as a word B such that the sentence "B is a (kind of) A" is acceptable. For example, *colour* is a hypernym of *red* as the sentence "red is a kind of colour" is true, and thus *red* is a hyponym of *colour*. This brings us to the second advantage of allowing nouns into our classification system; like adjectives, they can be ordered (this time in a tree[11]) by the hypernym-hyponym relationship.

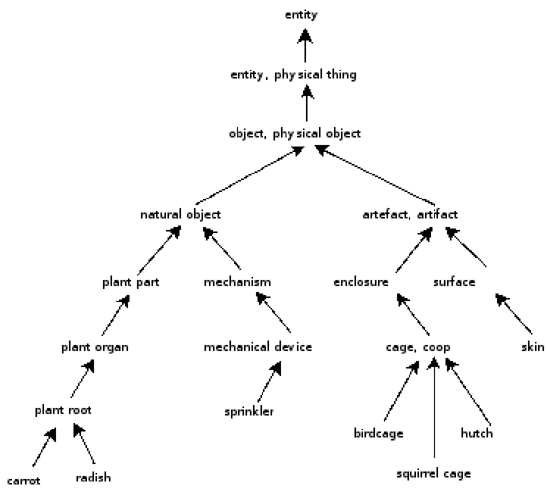


FIGURE 4. An example of a hypernym-hyponym tree from WordNet. Image: [25].

There is already a substantial amount of work done on classifying nouns by the hypernym-hyponym relationship, and there exist algorithms which extract this sort of structure from a given corpus [4, 16, 17, 26]. To elaborate further, Hearst [17] in 1992 revolutionarily algorithmitised hypernym-hyponym relationships according to a certain set of English rules (which, incidentally, can be recreated for Irish). Caraballo [4] took this work further and produced a working example with the 'Wall Street Journal' Penn Treebank corpus [23]. Gruenstein [16] produced a survey of the methods of (primarily) Hearst and Caraballo which gives a good explanation of the algorithms involved, too.

As well as this, there already exists the knowledge base WordNet [31] and its Irish counterpart LSG (Líonra Séimeantach na Gaeilge) [28], both of which have organised

[11] It might not be technically correct to refer to the structure as a tree, as each word might have several hypernyms. Nevertheless, the terminology has stuck.

thousands of nouns into this hierarchical relationship. Therefore we shall assume a hierarchy such as *food* $\to$ *fruit* $\to$ *berry* can already be extracted from text.

Using these tools, we have the following options when making use of the hypernym-hyponym tree in conceptual space creation:

(1) If we are interested solely in conceptual space creation (i.e. are only concerned with conceptual spaces for *one* language) we can remove the dependency of the tree on the corpus being analysed by using WordNet to create a hypernym-hyponym tree such as *Figure 4*. Relabelling the vertices gives us a convex space associated to each noun in the text via their path from root to leaf. (For an example of this, and more details, see *Section 6.1*).

(2) If we are interested in using conceptual spaces for language translation the matter becomes trickier - the trees generated by WordNet and LSG might not have the same structure. In personal communication with the author, Scannell [28] shared the source LSG files which confirm that, although linked with the Princeton English WordNet hence containing a similar structure, LSG is as of the time of writing not as well connected in hypernym-hyponym relationships as its English counterpart. We can recover from this as follows:

   (a) In one direction we could only use nouns already translated to describe a new noun. Take for example the Irish words úll - apple, and carr - car. If we knew the translations of torthaí - fruit, and feithicil - vehicle, by using the English WordNet tree we have a way of distinguishing between the nouns úll and carr purely numerically, through a labelled hypernym-hyponym tree.

   (b) If we assume we are given two copies of the same corpus, one in English and the other in Irish, then we can assume the *same* (up to synonyms, maybe) hierarchy of nouns is produced in the corresponding languages, using the extraction algorithms created by Hearst and Caraballo ([17], [4], resp.).

   (c) In other languages such as French [30] or Italian [19], the WordNet tree is more complete and more closely resembles the English WordNet tree. This should not be so surprising - In French, Italian and Irish, the English WordNet tree has been the starting point, and the main body of work comes from, in effect, translating the English WordNet tree to French, Italian, Irish, etc. This is a slow process, which is necessarily done by hand[12] however is on the way to being completed[13]. Therefore one day the LSG will be as rich and complicated as its English counterpart.
   With this in mind, it would be possible to simplify the English Wordnet tree in order for it to be directly comparable to a WordNet tree in another language.

The key point: given a corpus of text in English producing the hierarchy *food* $\to$ *fruit* $\to$ *berry*, we can assume the hierarchy *bia* $\to$ *torthaí* $\to$ *caora* produced by the Irish corpus is *directly comparable* to the English hierarchy, meaning we can instead label the hierarchy as $v_0 \to v_1 \to v_2$ and refer to berry (and caora) by its path in the hierarchy: $\{v_0, v_1, v_2\}$.

So when translating between two languages, we need not translate the nouns in our hypernym-hyponym tree. *Sections 6.1 & 6.2* contain examples of this proposal working successfully.

[12]For example, Irish distinguishes between dearg and rua. Two words which might be translated in English to *red*, however the latter is only ever used in describing people with red hair. Thus human translators are needed to initially make these distinctions.

[13]Noted in personal communication with the creator of the LSG [28].

5.3. **Verbs.** Provisionally, the power of verbs in our model isn't particularly strong. We can use (intransitive) verbs to group nouns based on the nouns' actions or how they are acted upon. From this, in each of our respective languages we can ascribe a subset of nouns to a given noun. For example, we may have difficulty distinguishing between apples and roses in some corpus, as they might both be closely related in terms of colour descriptors and smell. However we could distinguish *apple* and *rose* using the verb *eat*; apples are eaten, whereas roses are not, so if

$$N^{\text{eat}} := \{n : n \text{ is a noun which is eaten}\}, \qquad \text{apple} \in N^{\text{eat}}, \ \text{rose} \notin N^{\text{eat}}.$$

This suggestion ultimately falls beyond the extent of this essay; in *Section 6* and onward we will be discussing the applications and results of *Sections 5.1 & 5.2*.

6. Automatic Conceptual Space Creation from a Corpus

The first hurdle we must overcome if we wish to use the DisCoCat machinery is taking words in our foreign language (here Irish) and systematically representing them as convex spaces. The method we propose is reminiscent of how language is learnt in humans - if one tells you an úll is a red, round, smooth, bitter or sweet fruit, you will (eventually, with enough information) come to understand one is describing an apple. It is in this vein we present the following definition:

Definition 6.1. A *descriptor* D of a noun N is an adjective or noun which aids in the description of N; if D is an adjective it describes physical properties of N (e.g. *red, bitter, smooth*) and if D is a noun it classifies N according to nouns in an already-known hierarchical structure (e.g *fruit*, belonging to *food* $\to$ *fruit* $\to$ *berry*).

The necessity of adjectives as descriptors is immediate; after all an adjective is commonly defined as a word describing a noun. Defining other nouns to be descriptors might initially seem unnecessary, however it is clear they still carry information about the noun they are describing, as detailed in *Section 5.2*.

The basic idea of automatic conceptual space creation we propose is as follows: given a corpus of text involving heavy use of a noun N, parse the text identifying descriptors of N. Adjective descriptors can be given numerical values and represented in a high dimensional vector space according to *Section 5.1*. Taking the convex hull of the points in each adjective type, then the tensor product of the convex hulls, we represent the adjective descriptors of N as a convex set. Noun descriptors can be placed in a hierarchical tree and represented as a convex set a là *Section 5.2*. Combining these convex subsets under a tensor product once more gives us a conceptual space, as required.

6.1. **Example: Going Bananas.** Suppose we are given the following corpus of text:

Corpus 6.2. The banana, a fruit, looks long and yellow. Bananas can be mushy or just soft. After some time, bananas turn brown. Originally bananas are green. Bananas taste sweet but a little bitter. In some countries a banana is also a dessert. □

Let N = *banana*. The descriptors of N are the following:

Adjectives	long, yellow, mushy, soft, brown, green, sweet, a little bitter.
Nouns	fruit, dessert.

We first deal with the adjectives. We shall organise them according to *Section 5.1*. Define the noun spaces

$$N_{\text{dimension}} = Conv(\text{long}) = \{\text{long}\} = \{0.75\}.$$

$$N_{\text{colour}} = Conv(\text{yellow} \cup \text{green} \cup \text{brown}).$$

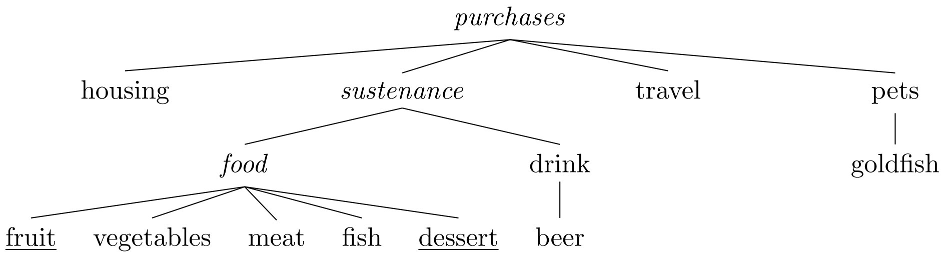
$$N_{\text{savour}} = Conv(\text{sweet} \cup \text{a little bitter}).$$
$$N_{\text{texture}} = Conv(\text{mushy} \cup \text{soft}) = [0.25, 0.5],$$

The adjective descriptor is defined as

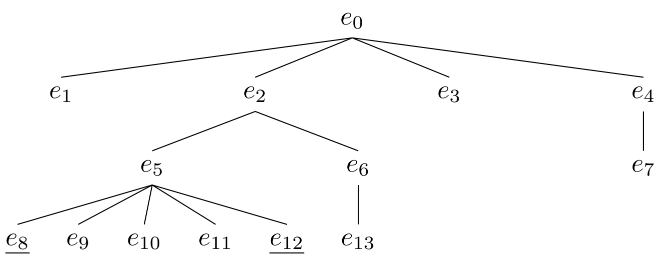
$$D_{\text{adj}} = N_{\text{dimension}} \otimes N_{\text{colour}} \otimes N_{\text{savour}} \otimes N_{\text{texture}}.$$

Note that if we are working off the list in *Section 5.1*, many adjective types have been skipped - **Age**, **Value**, **Speed**, etc. were not relevant. Thus, their corresponding noun spaces are empty. We will work off the assumption that the ordering of noun spaces is fixed by the list in *Section 5.1* and when formally writing and combining conceptual spaces we should include a symbol, e.g. $\emptyset_{\text{age}}$, to indicate the noun space N_{age} is present, just empty.

Next, consider the nouns. Imagine we have an already existing hierarchical structure in which the descriptor nouns $D_1, \dots D_n$ are already present (e.g. a hypernym-hyponym tree given by WordNet). In order to represent N as a convex set here, we take all direct ancestors of $D_1, \dots D_n$. In our example, *banana* is described by *fruit* and *dessert*. Suppose the following tree is created from WordNet[14]:



Therefore $D_{\text{noun}} := \{\text{purchases, sustenance, food, fruit, dessert}\}$. Let us label the above hypernym-hyponym tree as follows:



With this labelling[15] D_{noun} becomes $\{e_0, e_2, e_5, e_8, e_{12}\}$. The conceptual space for banana is then given by

$$\text{banana} := D_{\text{adj}} \otimes D_{\text{noun}}.$$

The advantage of this? Preforming the same algorithm on the **same** corpus of text, this time in Irish:

Corpus 6.3. Breathnaíonn an banana, torthaí, cosúil le fada agus buí. Is féidir le bananaí a bheith maothlach nó díreach bog. Tar éis roinnt ama, éiríonn bananaí donn. Ar dtús, tá bananaí glas. Blas bananaí milis ach beagán searbh. I roinnt tíortha is milseog é banana freisin. □

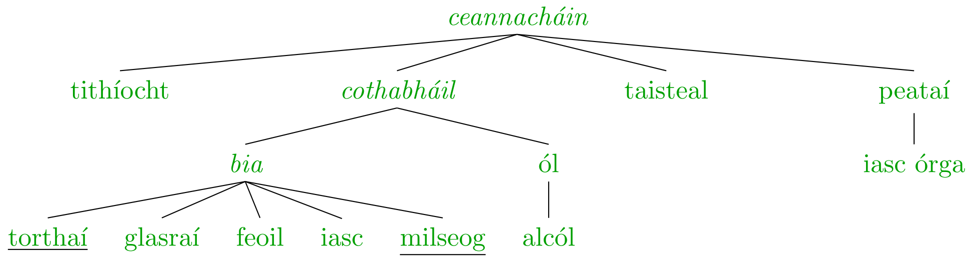
[14] It isn't, but makes for a more tangible example.

[15] Now independent of the English language.

...leads us to the following conceptual space definition for *banana*:

$$\text{banana} = Conv(\text{fada}) \otimes Conv(\text{buí} \cup \text{glas} \cup \text{donn}) \\ \otimes Conv(\text{milis} \cup \text{beagán searbh}) \otimes Conv(\text{maothlach} \cup \text{bog}) \otimes \{e_0, e_2, e_5, e_8, e_{12}\}$$

...a similar space to *banana* in English, assuming we also have the following tree:



Note that we are also making the assumption that all of the work of *Section 5* is done in Irish, too - any list of properties, say **Colour** or **Texture**, need to be manually entered in Irish as well as English. However this is only the case for the *adjectives* - as mentioned in *Section 5.2*, translations of the nouns need not be provided. Instead, the algorithm generating the hypernym-hyponym trees in English and Irish, or WordNet, does the work required.

This was a very simple, almost trivial example to get things going. In particular, in *Corpus 6.2* there was only one noun of interest; banana. In the following longer corpus there are multiple nouns with many descriptors, meaning when we attempt to translate Iúpatar in *Section 7* it will be a nontrivial exercise, requiring us to search through and compare our conceptual spaces.

6.2. **Another Example: Planets, the Sun and More Fruit.**

Corpus 6.4. Venus is a planet in the solar system. Venus has a solid and rocky surface. Venus is the second planet in the solar system and is called Earth's sister because it is nearly the same size as Earth. Venus is very hot and the pressure on its surface is high. Venus is bright in the night sky and looks like a ball.

Jupiter, another planet in the solar system, also looks like a ball. Jupiter sits in outer space. The size of Jupiter is very large; it is the largest planet in the solar system. Jupiter is called a gas giant because it is large and gassy. Jupiter is primarily orange and brown and red in colour. Jupiter is far away from Earth. It is very windy on Jupiter and also freezing cold. Jupiter is very bright in the night sky.

Mars is a planet next to Earth. Mars is coloured very red, and brown and orange. Mars is cold, but not very cold. Mars is smaller than Earth. Mars is rocky like Venus and Earth. Mars sits in outer space.

Apples are fruits. Apples are round and soft. Apples can be red or green, and they can be eaten for dessert. Some apples taste bitter and other apples taste sweet. An apple looks like a ball.

The Sun is a star, not a planet. It sits in the centre of the solar system. The Sun is the brightest thing in the sky. The Sun is huge and very hot. The Sun is round and also looks like a ball. The gravity on the Sun is very strong, meaning it is very dense. □

Let us examine five main nouns from this corpus;

$$N^1 = \text{Venus},\ N^2 = \text{Jupiter},\ N^3 = \text{Mars},\ N^4 = \text{apple},\ N^5 = \text{The Sun}.$$

Organising this into a table we obtain:

Venus	**Adjectives**	solid, rocky, same size as Earth, hot, high pressure, bright.
	Nouns	planet, Earth's sister, ball.
Jupiter	**Adjectives**	very large, gassy, orange, brown, red, far away, windy, freezing, very bright.
	Nouns	planet, outer space, ball.
Mars	**Adjectives**	very red, brown, orange, cold, smaller than Earth, rocky.
	Nouns	planet, outer space.
Apple	**Adjectives**	round, soft, red, green, bitter, sweet.
	Nouns	fruit, ball.
The Sun	**Adjectives**	brightest, huge, very hot, round, very dense.
	Nouns	star, ball.

We first deal with the adjectives. These can be organised according to *Section 5.1*:

(1) **Venus.**

$$\begin{aligned}
N_{\text{dimension}} &= Conv(\text{same size as Earth}) = \{0.5\},\\
N_{\text{intensity}} &= Conv(\text{bright}) = \{0.7\},\\
N_{\text{temperature}} &= Conv(\text{hot}) = \{0.75\},\\
N_{\text{density}} &= Conv(\text{solid}) = \{0.9\},\\
N_{\text{texture}} &= Conv(\text{rocky}) = \{0.9\}.
\end{aligned}$$

D^1_{adj} is the tensor product of these. Note that we were required to drop some adjectives, such as *high pressure*, as our adjective classification from *Section 5.1* is not specific enough to capture all details.

(2) **Jupiter.**

$$\begin{aligned}
N_{\text{dimension}} &= Conv(\text{very large}) = \{0.7\},\\
N_{\text{colour}} &= Conv(\text{orange} \cup \text{brown} \cup \text{red}),\\
N_{\text{intensity}} &= Conv(\text{very bright}) = \{0.8\},\\
N_{\text{temperature}} &= Conv(\text{freezing}) = \{0\},\\
N_{\text{density}} &= Conv(\text{gassy}) = \{0.1\}.
\end{aligned}$$

D^2_{adj} is the tensor product of these.

(3) **Mars.**

$$\begin{aligned}
N_{\text{dimension}} &= Conv(\text{smaller than Earth}) = \{0.25\},\\
N_{\text{colour}} &= Conv(\text{red} \cup \text{brown} \cup \text{orange}),\\
N_{\text{temperature}} &= Conv(\text{cold}) = \{0.4\},\\
N_{\text{texture}} &= Conv(\text{rocky}) = \{0.9\}.
\end{aligned}$$

D^3_{adj} is the tensor product of these. Recall that if we want to write the tensor product completely correct and formally, we must also include symbols $\emptyset_{\text{age}}$, $\emptyset_{\text{value}}$, $\emptyset_{\text{smell}}$, $\emptyset_{\text{savour}}$, $\emptyset_{\text{sound}}$, $\emptyset_{\text{density}}$, $\emptyset_{\text{mass}}$, $\emptyset_{\text{speed}}$ in the appropriate places.

(4) **Apple.**

$$N_{\text{colour}} = Conv(\text{red} \cup \text{green}),$$
$$N_{\text{taste}} = Conv(\text{bitter} \cup \text{sweet}),$$
$$N_{\text{texture}} = Conv(\text{soft}) = \{0.4\}.$$

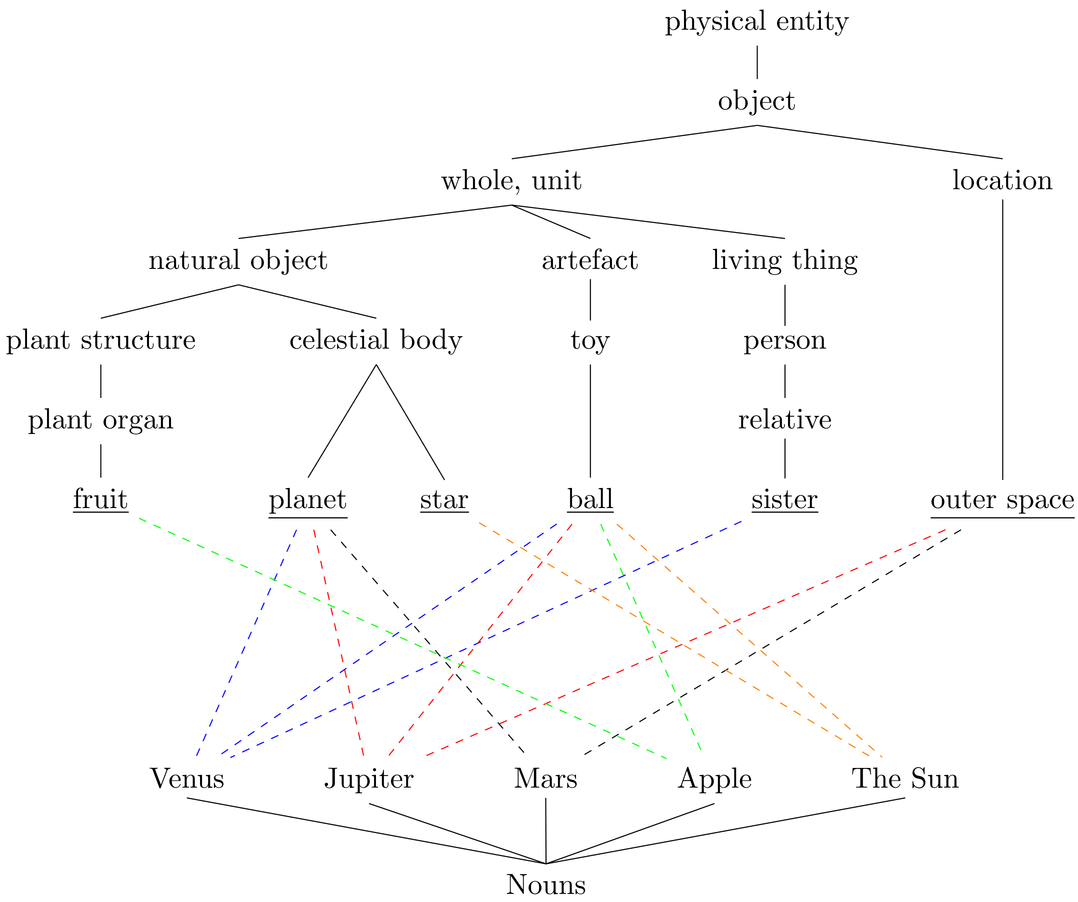
Once again D^3_{adj}, the tensor product of these.

(5) **The Sun.**

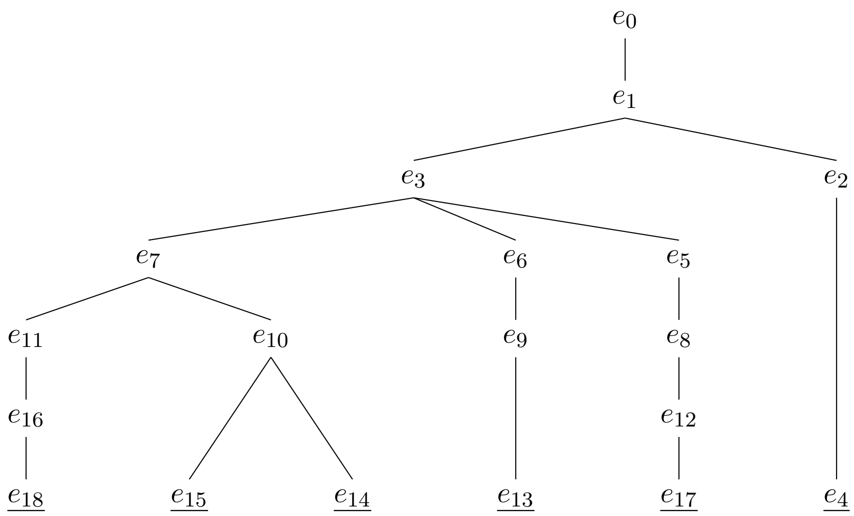
$$N_{\text{dimension}} = Conv(\text{huge}) = \{1\},$$
$$N_{\text{intensity}} = Conv(\text{brightest}) = \{1\},$$
$$N_{\text{temperature}} = Conv(\text{very hot}) = \{1\},$$
$$N_{\text{density}} = Conv(\text{very dense}) = \{1\}.$$

Finally D^5_{adj} is the tensor product of these.

At this point we still might not have an accurate picture of the situation. For instance, the Sun, although it is an astronomical body it does not seem to share many things in common with the planets - no colour or texture has been given for it. Also, painting an apple as a "red or green, bitter or sweet, soft thing" doesn't create the same conceptual space as when we mention the fact an apple is a fruit. This additional information is added by the following tree, generated by WordNet [31]:



Relabel the nodes of the tree as follows:



Then we can define:

$$
\begin{aligned}
D^1_{\text{noun}} &= \{e_0, e_1, e_3, e_5, e_6, e_7, e_8, e_9, e_{10}, e_{12}, e_{13}, e_{15}, e_{17}\},\\
D^2_{\text{noun}} &= \{e_0, e_1, e_2, e_3, e_4, e_6, e_7, e_9, e_{10}, e_{13}, e_{15}\},\\
D^3_{\text{noun}} &= \{e_0, e_1, e_2, e_3, e_4, e_7, e_{10}, e_{15}\},\\
D^4_{\text{noun}} &= \{e_0, e_1, e_3, e_6, e_7, e_9, e_{11}, e_{13}, e_{16}, e_{18}\},\\
D^5_{\text{noun}} &= \{e_0, e_1, e_3, e_6, e_7, e_9, e_{10}, e_{13}, e_{14}\},
\end{aligned}
$$

and finally we obtain the conceptual spaces

$$
\begin{aligned}
&\text{Venus} := D^1_{\text{adj}} \otimes D^1_{\text{noun}},\\
&\text{Jupiter} := D^2_{\text{adj}} \otimes D^2_{\text{noun}},\\
&\text{Mars} := D^3_{\text{adj}} \otimes D^3_{\text{noun}},\\
&\text{Apple} := D^4_{\text{adj}} \otimes D^4_{\text{noun}},\\
&\text{The Sun} := D^5_{\text{adj}} \otimes D^5_{\text{noun}}.
\end{aligned}
$$

What has this captured? The data for *Jupiter* tells us this noun is a relatively large, orange, red & brown, quite bright, freezing cold, low density planet; which is a celestial body, natural object (...) and also a ball; a type of toy, an artefact (...) and it sits in outer space; which is a location (...). Note that our account isn't entirely accurate - Jupiter is not a toy, for instance[16]. However, the beauty of DisCoCat models is the objective truth of the statement does not matter, rather the truth of the statement in context. So while a description as 'toy' might not be empirically accurate for Jupiter, *Corpus 6.4* describes Jupiter, the Sun and apples all as *balls*, hence relative to this corpus it is fitting they are all described in an equal manner such as 'toy'. Of course, the reason humans would not categorise Jupiter as a toy is we understand what a *toy* is - this algorithm does not; it uses toy as a method of grouping certain nouns together based on the corpus. In conclusion: describing Jupiter as a toy might seem odd for a human's conceptual space, but for a machine's it is a meaningless word used to connect two nouns it is attempting to understand.

In Irish, the same corpus is as follows:

[16]This descriptor arose as we used a simile in describing Jupiter.

Corpus 6.5. Is í Véineas plánéad sa ghrianchóras. Tá dromchla tathagach agus carraigeach ag Véineas. Is í Véineas an dara plánéad sa ghrianchóras agus glaotar deirfiúr an Domhan í mar tá sí beagnach an méid céanna leis an Domhan. Tá sé an-te ar Véineas agus tá an brú ar a dromchla ard. Tá Véineas geal i spéir na hoíche agus breathnaíonn sí cosúil le liathróid.

Breathnaíonn Iúpatar, plánéad eile sa ghrianchóras, cosúil le liathróid freisin. Suíonn Iúpatar i spás seachtrach. Tá Iúpatar an-mhór; tá sé an plánéad is mó sa ghrianchóras. Fatach gáis a ghlaotar ar Iúpatar mar tá sé mór agus déanta as gáis. Tá Iúpatar go príomha oráiste agus donn agus dearg i ndath. Tá Iúpatar i bhfad gcéin ó an Domhan. Tá sé an-ghaothmhar ar Iúpatar agus an-fhuar freisin. Tá Iúpatar an-gheal i spéir na hoíche.

Is é Mars pláinéid in aice leis an Domhan. Tá Mars daite an-dearg, agus donn agus oráiste. Tá sé fuar ar Mars, ach níl sé an-fhuar. Tá Mars níos lú ná an Domhan. Tá Mars carraigeach cosúil le Véineas agus an Domhan. Suíonn Mars i spás seachtrach.

Is torthaí iad úlla. Tá úlla liathróideach agus bog. Féadfaidh úlla a bheith dearg nó glas, agus is féidir iad a ithe mar milseog. Tá blas searbh ar roinnt úill agus blas milis ar úlla eile. Breathnaíonn úll cosúil le liathróid.

Is réalta í an grian, ní phláinéid. Tá sí suite i lár an chórais ghréine. Is í an grian an rud is gile sa spéir. Tá an grian ollmhór agus an-te. Tá an grian liathróideach agus breathnaíonn sí ar liathróid fresin. Tá an imtharraingt ar an ghrian an-láidir, rud a chiallaíonn go bhfuil sí an-dlúth. □

The five main nouns of this corpus are (in no particular order)

$$M^1 = \text{Véineas},\ M^2 = \text{Iúpatar},\ M^3 = \text{Mars},\ M^4 = \text{Úll},\ M^5 = \text{Grian}.$$

Organising the information of *Corpus 6.5* into a table we obtain:

Véineas	**Adjectives**	tathagach, carraigeach, beagnach an méid céanna leis an Domhan, an-te, brú . . . ard, geal.	solid, rocky, same size as Earth, hot, high pressure, bright.
	Nouns	plánéad, deirfiúr an Domhan, liathróid.	planet, Earth's sister, ball.
Iúpatar	**Adjectives**	an-mhór, déanta as gáis, oráiste, donn, dearg, i bhfad i gcéin, an-ghaothmhar, an-fhuar, an-gheal.	very large, gassy, orange, brown, red, far away, windy, freezing, very bright.
	Nouns	plánéad, spás seachtrach, liathróid.	planet, outer space, ball.
Mars	**Adjectives**	an-dearg, oráiste, donn, fuar, níos lú ná an Domhan, carraigeach.	very red, brown, orange, cold, smaller than Earth, rocky.
	Nouns	plánéad, spás seachtrach.	planet, outer space.
Úll	**Adjectives**	liathróideach, bog, dearg, glás, searbh, milis.	round, soft, red, green, bitter, sweet.
	Nouns	torthaí, liathróid.	fruit, ball.
Grian	**Adjectives**	an rud is gile, ollmhór, an-te, liathróideach, an-dlúth.	brightest, huge, very hot, round, very dense.
	Nouns	réalta, liathróid.	star, ball.

We first deal with the adjectives. These can be organised according to *Section 5.1*:

(1) **Véineas.**

$$N_{\text{dimension}} = Conv(\text{beagnach an méid céanna leis an Domhan}) = \{0.5\},$$
$$N_{\text{intensity}} = Conv(\text{geal}) = \{0.6\},$$
$$N_{\text{temperature}} = Conv(\text{an-te}) = \{0.85\},$$
$$N_{\text{density}} = Conv(\text{tathagach}) = \{0.9\},$$
$$N_{\text{texture}} = Conv(\text{carraigeach}) = \{0.9\}.$$

Note that the values here are different than the corresponding values in English for geal (bright), an-te (hot), etc. For example, in Irish there is no word for "hot" - to describe high temperatures there is just "warm" and "very warm"[17]. So "an-te" (literally translated as "very warm") suffices for "hot", therefore since "an-te" is the hottest the weather can be described, it is assigned a value of 0.85 in Irish (because in English, "very hot" would need to correspond to a higher value than "hot", which is 0.75).

$\overline{D}^1_{\text{adj}}$ is the tensor product of $N_{\text{dimension}}, \dots, N_{\text{texture}}$. Note that we were required to drop some adjectives, such as *brú... ard* (*high pressure*), as our adjective classification from *Section 5.1* is not specific enough to capture all details.

(2) **Iúpatar.**

$$N_{\text{dimension}} = Conv(\text{an-mhór}) = \{0.8\},$$
$$N_{\text{colour}} = Conv(\text{oráiste} \cup \text{donn} \cup \text{dearg}),$$
$$N_{\text{intensity}} = Conv(\text{an-geal}) = \{0.7\},$$
$$N_{\text{temperature}} = Conv(\text{an-fuar}) = \{0.1\},$$
$$N_{\text{density}} = Conv(\text{déanta as gáis}) = \{0.1\}.$$

$\overline{D}^2_{\text{adj}}$ is the tensor product of these. Again, we loose information such as an-ghaothmhar (very windy) as we cannot yet capture all adjectives with our algorithm.

(3) **Mars.**

$$N_{\text{dimension}} = Conv(\text{níos lú ná an Domhan}) = \{0.25\},$$
$$N_{\text{colour}} = Conv(\text{dearg} \cup \text{donn} \cup \text{oráiste}),$$
$$N_{\text{temperature}} = Conv(\text{fuar}) = \{0.4\},$$
$$N_{\text{texture}} = Conv(\text{carraigeach}) = \{0.9\}.$$

$\overline{D}^3_{\text{adj}}$ is the tensor product of these. Recall that is we want to write the tensor product completely correct and formally, we must also include symbols $\emptyset_{\text{age}}$, $\emptyset_{\text{value}}$, $\emptyset_{\text{smell}}$, $\emptyset_{\text{savour}}$, $\emptyset_{\text{sound}}$, $\emptyset_{\text{density}}$, $\emptyset_{\text{mass}}$, $\emptyset_{\text{speed}}$ in the appropriate places.

(4) **Úll.**

$$N_{\text{colour}} = Conv(\text{dearg} \cup \text{glás}),$$
$$N_{\text{taste}} = Conv(\text{searbh} \cup \text{milis}),$$
$$N_{\text{texture}} = Conv(\text{bog}) = \{0.4\}.$$

[17] In Ireland, one does not encounter high temperatures often enough to justify another word.

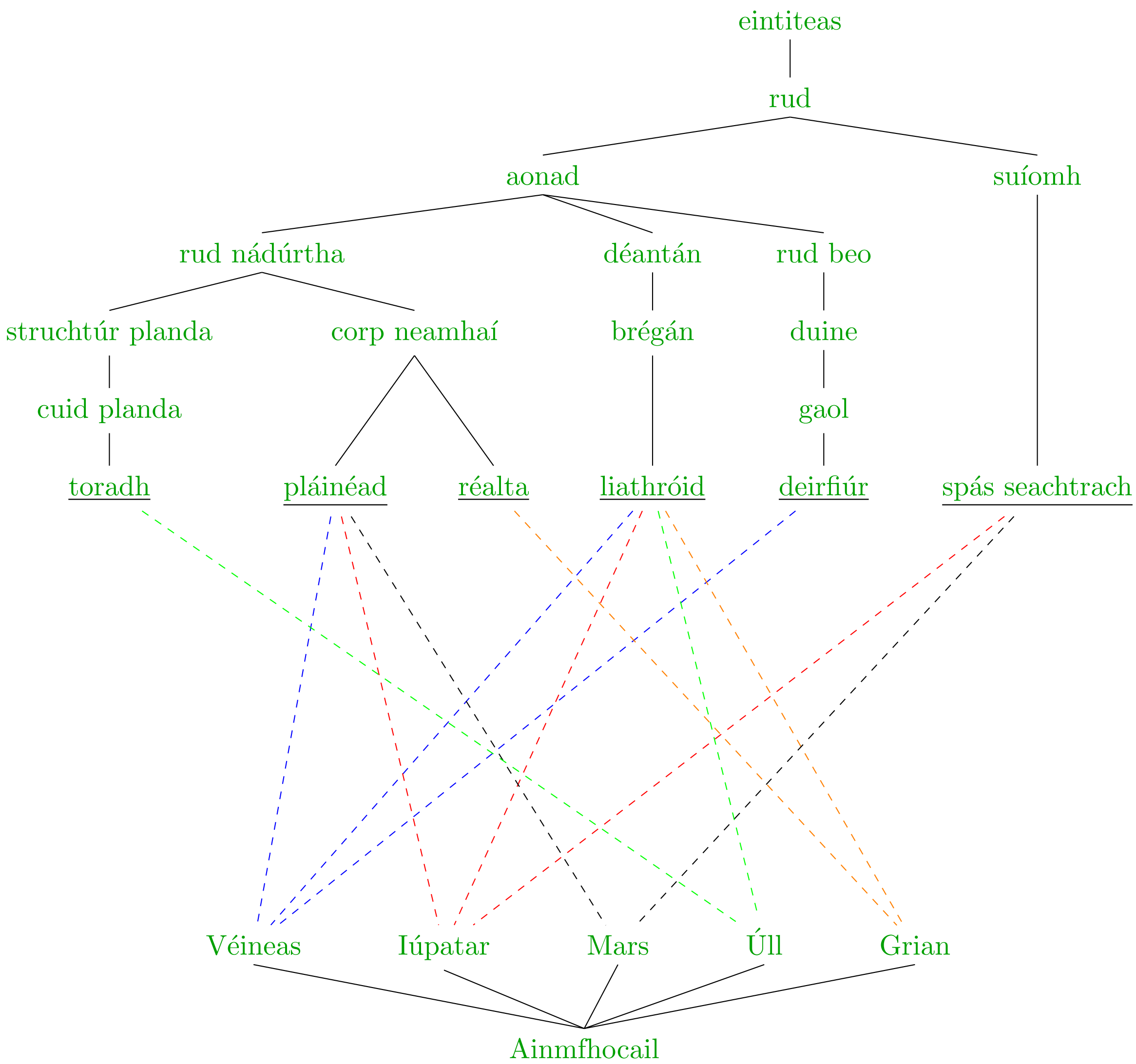
Once again $\overline{D}^3_{\text{adj}}$, the tensor product of these.

(5) **Grian.**

$$N_{\text{dimension}} = Conv(\text{ollmhór}) = \{0.9\},$$
$$N_{\text{intensity}} = Conv(\text{an rud is gile}) = \{1\},$$
$$N_{\text{temperature}} = Conv(\text{an-te}) = \{0.85\},$$
$$N_{\text{density}} = Conv(\text{an-dlúth}) = \{1\}.$$

Finally $\overline{D}^5_{\text{adj}}$ is the tensor product of these.

The additional linguistic information from the descriptor nouns is obtained by referencing a hypernym-hyponym tree, which e.g. WordNet in Irish organises as:



If we relabel the tree as:

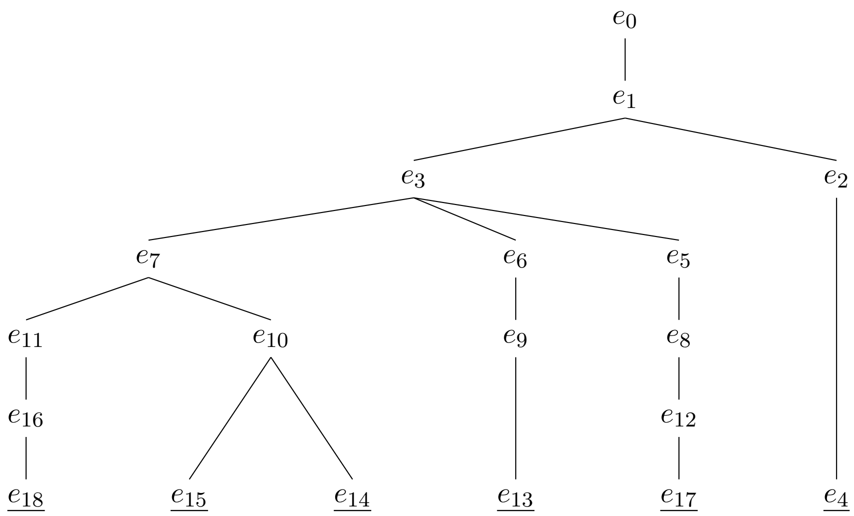


FIGURE 5. Hypernym-Hyponym tree for *Corpora 6.4 & 6.5.*

...we can define:

$$\begin{aligned}
\overline{D}^1_{\text{noun}} &= \{e_0, e_1, e_3, e_5, e_6, e_7, e_8, e_9, e_{10}, e_{12}, e_{13}, e_{15}, e_{17}\},\\
\overline{D}^2_{\text{noun}} &= \{e_0, e_1, e_2, e_3, e_4, e_6, e_7, e_9, e_{10}, e_{13}, e_{15}\},\\
\overline{D}^3_{\text{noun}} &= \{e_0, e_1, e_2, e_3, e_4, e_7, e_{10}, e_{15}\},\\
\overline{D}^4_{\text{noun}} &= \{e_0, e_1, e_3, e_6, e_7, e_9, e_{11}, e_{13}, e_{16}, e_{18}\},\\
\overline{D}^5_{\text{noun}} &= \{e_0, e_1, e_3, e_6, e_7, e_9, e_{10}, e_{13}, e_{14}\},
\end{aligned}$$

and finally we obtain the conceptual spaces

$$\begin{aligned}
\text{Véineas} &:= \overline{D}^1_{\text{adj}} \otimes \overline{D}^1_{\text{noun}},\\
\text{Iúpatar} &:= \overline{D}^2_{\text{adj}} \otimes \overline{D}^2_{\text{noun}},\\
\text{Mars} &:= \overline{D}^3_{\text{adj}} \otimes \overline{D}^3_{\text{noun}},\\
\text{Úll} &:= \overline{D}^4_{\text{adj}} \otimes \overline{D}^4_{\text{noun}},\\
\text{Grian} &:= \overline{D}^5_{\text{adj}} \otimes \overline{D}^5_{\text{noun}}.
\end{aligned}$$

as desired.

7. Sentence Meaning and the Category **ConvexRel**

In 2004 Gärdenfors [10, 11, 12] introduced conceptual spaces as a means of representing information in a 'human' way; the founding idea being if two objects represent the same concept, then every object somehow 'in between' these objects also represents the same concept. We can mathematically describe the property of 'in between' via *convex algebras*, an introduction to which is given by Bolt et al. [3, §4].

For a set X, let $D(X)$ be the set of all finite formal sums $\sum_i p_i|x_i\rangle$ where $x_i \in X$, $p_i \in \mathbb{R}^{\geq 0}$ and $\sum_i p_i = 1$. Define:

Definition 7.1. A **convex algebra** is a set A with a function $\alpha : D(A) \to A$ known as a *mixing operation* such that

- $\alpha(|a\rangle) = a$,
- $\alpha\left(\sum_{i,j} p_i q_{ij} |a_{ij}\rangle\right) = \alpha\left(\sum_i p_i \middle| \alpha\big(\sum_j q_{ij} |a_{ij}\rangle\big) \right\rangle\Big)$.

The two convex algebras of interest to us are *Examples 9 & 14* of [3].

(1) The closed real interval $[0, 1]$ has a convex algebra structure induced by the vector space $\mathbb{R}$. The formal sums $\sum_i p_i |x_i\rangle$ are sums of elements in $[0, 1]$ with addition and multiplication from $\mathbb{R}$. The mixing operation is the identity map.

(2) A finite tree can be a convex algebra - in particular, the hypernym-hyponym trees we are interested in are affine semilattices, hence the formal sums

$$\sum_i p_i |a_i\rangle := \bigvee_i \{a_i : p_i > 0\}$$

are well defined. (So, for example the formal sum $p_1|x_1\rangle + p_2|x_2\rangle + p_3|x_3\rangle$ is the lowest level in the tree containing x_1, x_2, x_3; their *join*.)

In order to identify the category **ConvexRel** we also need to define *convex relations*:

Definition 7.2. Let A, B be sets with mixing operations α, β respectively. A **convex relation** $(A, \alpha) \to (B, \beta)$ is a binary relation $R \subseteq A \times B$ (also written $R : A \to B$) that respects forming mixtures:

$$\forall i \; R(a_i, b_i) \quad \Rightarrow \quad R\left(\alpha\big(\sum_i p_i |a_i\rangle\big), \beta\big(\sum_i q_i |b_i\rangle\big) \right).$$

ConvexRel is a category with convex algebras as objects and convex relations as morphisms. It is compact closed [3, Theorem 1], hence (by Coecke et al. [6]) combines perfectly with the Lambek grammar category allowing us to create a morphism to interpret meanings in the **ConvexRel** category via the type reductions in the Lambek grammar category. Since the conceptual spaces created by methods from *Sections 5 & 6* are members of the **ConvexRel** category we can use Lambek grammar rules for Irish and English to compare and calculate the meanings of sentences.

The work of *Sections 5 & 6* are solely concerned with conceptual spaces for nouns. Gärdenfors [12] has written about verb spaces, adjective spaces, and other spaces for parts of speech, and Bolt et al. [3, §5.1.2-5.1.3] have produced examples of simple, hand crafted conceptual spaces for adjective and verbs, but it is beyond the scope of this paper to algorithmically create conceptual spaces for linguistic structures other than nouns.

7.1. **Metrics for Conceptual Spaces.** Our final goal is to compare the conceptual spaces created in *Section 6.2* in Irish and English. To do this we require some measure of distance between concepts; we require a metric on **ConvexRel**. First let us introduce the following notions from [24]:

Definition 7.3. A *quantale* is a join complete partial order Q with a monoid structure $(\otimes, k)$ satisfying the following distributivity axioms:

$$\text{For all } a, b \in Q \text{ and } A, B \subseteq Q,$$
$$a \otimes \left[\bigvee B\right] = \bigvee \{a \otimes b : b \in B\},$$
$$\left[\bigvee A\right] \otimes b = \bigvee \{a \otimes b : a \in A\}.$$

Moreover, a quantale is said to be *commutative* if its monoid structure is commutative.

Example 7.4. The *Lawvere quantale* C is a commutative quantale whose underlying set is the extended positive reals, written $[0,\infty]$, with reverse order and algebraic structure

$$\begin{aligned} \bigvee A &= \inf A, \\ a_1 \otimes a_2 &= a_1 + a_2, \\ k &= 0. \end{aligned}$$

$\Diamond$

One can think of a quantale Q as a "generalised truth space"; if a binary relation is described by its characteristic function $A \times B \to 2$, then a generalised binary relation is described by a characteristic function $A \times B \to Q$. In fact the binary relations of this appearance form a category $\mathbf{Rel}(Q)$. As mentioned by Marsden and Genovase [24], $\mathbf{Rel}(C)$ is a dagger compact closed category. This is in turn related to metrics, as the internal monads of $\mathbf{Rel}(C)$ - relations R satisfying

$$R(a,a) = 0 \quad \text{ and } \quad R(a,b) + R(b,c) \geq R(a,c)$$

- are *generalised metrics*, a term explained by Coecke et al. in [7].

Therefore if we consider $\mathbf{Rel}_{\mathtt{Convex}}(C)$, the category of C-relations with algebraic signature $\mathtt{Convex}$[18] then the internal monads are distance measures $d : A \times A \to [0,\infty]$ such that

$$d(a,a) = 0, \tag{D1}$$

$$d(a,b) + d(b,c) \geq d(a,c), \tag{D2}$$

$$d(a_1,a_2) + d(b_1,b_2) \geq d(pa_1 + (1-p)b_1, pa_2 + (1-p)b_2) \quad \text{for } p \in (0,1), \tag{D3}$$

according to [24, Example 7]. Thus if we consider the generalised 'taxicab' metric of $\mathbb{R}^n$:

$$d_t(a,b) = \sum_{i=1}^{n} |a_i - b_i|,$$

d_t is an example of such an internal monad. Also the 'path distance' metric on an affine semilattice T, given by

$$\begin{aligned} \text{for } p_1, p_2 \text{ paths in } T, \quad d_p(p_1,p_2) &= \max\{\#\text{nodes } p_1 \setminus p_2,\, \#\text{nodes } p_2 \setminus p_1\} \\ &= \text{“\# nodes } p_1 \text{ and } p_2 \text{ do not have in common”}, \end{aligned}$$

is also an internal monad of $\mathbf{Rel}_{\mathtt{Convex}}(C)$. (The properties (D1) and (D2) are straightforward to verify, and (D3) follows once we recall from [3, Example 13] that $\sum_i p_i |a_i\rangle = \bigvee_i \{a_i : p_i > 0\}$, hence is independant of the p_i.)

As the sum of two metrics is a metric, define the metric d on the conceptual spaces $D_{\text{adj}} \otimes D_{\text{noun}}$ we created in *Section 6*:

$$d\left(D^1_{\text{adj}} \otimes D^1_{\text{noun}}, D^2_{\text{adj}} \otimes D^2_{\text{noun}}\right) := \left(\sum_{\text{noun spaces } N_k \text{ of } D_{\text{adj}}} d_t(N^1_k, N^2_k) \right) + d_t(D^1_{\text{noun}}, D^2_{\text{noun}}),$$

where $d_t(N^1_k, N^2_k)$ is an extension of a metric to measure distances between sets:

$$d_t(N^1_k, N^2_k) = \begin{cases} \inf\{d_t(n_1,n_2) : n_1 \in N^1_k, n_2 \in N^2_k\} & \text{if } N^1_k \neq \emptyset, N^2_k \neq \emptyset, \\ \inf\{d_t(n_1,0) : n_1 \in N^1_k\} & \text{if } N^1_k \neq \emptyset, N^2_k = \emptyset, \\ \inf\{d_t(0,n_2) : n_2 \in N^2_k\} & \text{if } N^2_k \neq \emptyset, N^1_k = \emptyset, \\ 0 & \text{if } N^1_k = N^2_k = \emptyset, \end{cases}$$

and $D^1_{\text{noun}}, D^2_{\text{noun}}$ are paths in the hypernym-hyponym tree constructed from the corpus.

[18] $\mathbf{ConvexRel} = \mathbf{Rel}_{\mathtt{Convex}}(2)$.

Example 7.5. Consider the distance between "Apple" and "Jupiter", whose conceptual spaces were calculated in *Section 6.2*.

$$\begin{aligned}
d(\text{"Apple"}, \text{"Jupiter"}) &= \left(\sum_{\text{noun spaces } N_k \text{ of } D_{\text{adj}}} d_t(N_k^{\text{apple}}, N_k^{\text{jupiter}}) \right) + d_p(D_{\text{noun}}^{\text{apple}}, D_{\text{noun}}^{\text{jupiter}}) \\
&= \big(d_t(\emptyset, \{0.7\}) + d_t(Conv(\text{red} \cup \text{green}), Conv(\text{red} \cup \text{brown} \cup \text{orange})) \\
&\quad + d_t(\emptyset, \{0.8\}) + d_t(\emptyset, \{0\}) + d_t(Conv(\text{bitter} \cup \text{sweet}), \emptyset) + d_t(\emptyset, \{0.1\}) \\
&\quad + d_t(\{0.4\}, \emptyset)\big) + d_p(D_{\text{noun}}^{\text{apple}}, D_{\text{noun}}^{\text{jupiter}}) \\
&= (0.7 + 0 + 0.8 + \tfrac{\sqrt{3}}{3} + 0.1 + 0.4) + 4 \\
&= 6.577.
\end{aligned}$$

The calculation $d_t(Conv(\text{bitter} \cup \text{sweet}), \emptyset) = \frac{\sqrt{3}}{3}$ is excluded for brevity, but follows from calculations on Gärdenfors' taste tetrahedron (*Section 5.1*).

Note that one problem with defining "$d(N, \emptyset) = \inf\{d(n, 0) : n \in N\}$" is apparent in this example; $d_t(\emptyset, \{0\}) = 0$ but this is only because we haven't assigned a temperature to apples in *Corpus 6.4*. We do not usually picture apples as "freezing", hence in a more detailed corpus it would be the case $d_t(N_{\text{temperature}}^{\text{apples}}, N_{\text{temperature}}^{\text{jupiter}}) > 0$. However, we can only calculate with what is given to us in *Corpus 6.4*.

Similarly,

$$\begin{aligned}
d(\text{"Mars"}, \text{"Jupiter"}) &= 5.65, \\
d(\text{"Jupiter"}, \text{"Sun"}) &= 6.4901, \\
d(\text{"Apple"}, \text{"Sun"}) &= 8.977.
\end{aligned} \tag{17}$$

This seems to capture the rough picture we desire: conceptually, the planets Mars and Jupiter are close, while nouns like "Apple" and "Jupiter" or "Apple" and "Sun" are distant. "Sun" is also closer to "Jupiter" than to "Apple", as we might expect. ♢

Finally, let us return to translation between Irish and English.

Example 7.6. The distance between "Apple" and its Irish translation, "Úll", is given by

$$\begin{aligned}
d(\text{"Apple"}, \text{"Úll"}) &= \left(\sum_{\text{noun spaces } N_k \text{ of } D_{\text{adj}}} d_t(N_k^{\text{apple}}, N_k^{\text{úll}}) \right) + d_p(D_{\text{noun}}^{\text{apple}}, D_{\text{noun}}^{\text{úll}}) \\
&= (0 + 0 + 0) + 0 = 0,
\end{aligned}$$

which is to say as conceptual spaces, "Apple" and "Úll" are equal (as we might hope for a translation). On the other hand, the distance between "Apple" and "Grian" (English: "Sun") is

$$\begin{aligned}
d(\text{"Apple"}, \text{"Grian"}) &= \left(\sum_{\text{noun spaces } N_k \text{ of } D_{\text{adj}}} d_t(N_k^{\text{apple}}, N_k^{\text{grian}}) \right) + d_p(D_{\text{noun}}^{\text{apple}}, D_{\text{noun}}^{\text{grian}}) \\
&= (1 + \tfrac{\sqrt{3}}{3} + 0.4 + 0.9 + 1 + 0.85 + 1) + 3 \\
&= 8.727.
\end{aligned}$$

This seems like a fantastic result, however (like the distance between "Sun" and "Apple" in English) this calculation takes advantage of the fact that "Apple" and "Grian" (or "Apple" and "Sun") have no adjective descriptors in common. So, although it is an accurate and unsurprising result, we are in some sense 'lucky' *Corpora 6.4 & 6.5* did not highlight the similarities between apples and the Sun.

Pushing forward, we see

$$\begin{aligned} d(\text{“Sun”}, \text{“Grian”}) &= \left(\sum_{\text{noun spaces } N_k \text{ of } D_{\text{adj}}} d_t(N_k^{\text{sun}}, N_k^{\text{grian}}) \right) + d_p(D_{\text{noun}}^{\text{sun}}, D_{\text{noun}}^{\text{grian}}) \\ &= (0.1 + 0 + 0.15 + 0) + 0 \\ &= 0.25. \end{aligned}$$

Even though this is an exact translation, as conceptual spaces they are close but nonequal. This stems from the fact that adjectives can have different meanings with different intensities in different languages.

Finally, note that

$$\begin{aligned} d(\text{“Mars”}, \text{“Iúpatar”}) &= \left(\sum_{\text{noun spaces } N_k \text{ of } D_{\text{adj}}} d_t(N_k^{\text{mars}}, N_k^{\text{Iúpatar}}) \right) + d_p(D_{\text{noun}}^{\text{mars}}, D_{\text{noun}}^{\text{Iúpatar}}) \\ &= (0.55 + 0 + 0.7 + 0.3 + 0.1 + 0.9) + 3 \\ &= 5.55, \end{aligned}$$

so in Irish the conceptual spaces of Mars and Iúpatar are slightly closer than the corresponding spaces for Mars and Jupiter (cf. (17)). ♢

Finally, as promised at the end of *Section 6.1*, if we were to attempt to translate "Iúpatar" using the metric on **ConvexRel**, we see

$$\begin{aligned} d(\text{“Venus”}, \text{“Iúpatar”}) &= 7.5401, \\ d(\text{“Jupiter”}, \text{“Iúpatar”}) &= 0.3, \\ d(\text{“Mars”}, \text{“Iúpatar”}) &= 5.55, \\ d(\text{“Apple”}, \text{“Iúpatar”}) &= 6.6773, \\ d(\text{“Sun”}, \text{“Iúpatar”}) &= 6.1901. \end{aligned}$$

Hence choosing the conceptual space closest to "Iúpatar", which is "Jupiter", we deduce we have indeed successfully translated this word.

Remark 7.7. The beauty of attempting to translate by this method is we are comparing conceptual spaces built from individual corpora - no further knowledge of the word "Iúpatar" needs to be known in order to complete this exercise, and no other translations needed to be preformed beforehand! ♢

Remark 7.8. The author will admit this approach initially lacks the smoothness and cleanness of the vector space approach in *Sections 3 & 4* - for instance, in order for this approach to work in general it it necessary in both Irish and English to manually input values for the seven core adjective types (Dimension, Age, Colour, etc.). It is the opinion of the author, however, that such an exercise is an important one. This method is how we first master colours and smells and sizes; by hearing about them and memorising terms, ordered relative to each other. In the words of Gärdenfors [12], "we are not born with our concepts; they must be learned".

The author believes it is also necessary to preform this exercise separately for Irish, as adjectives in this language can have different emphases and occasionally different meanings! For example, in Irish there is a distinct between "dearg" and "rua". Both are translated as "red", however the latter is only ever used in describing a red-headed person. Thus the RGB values for "dearg" and "rua" are different for an Irish speaker, and the convex space model of meaning should reflect this. ♢

In conclusion, this essay has outlined two methods of translating from Irish to English using the distributional compositional categorical model of meaning; via vector spaces and the category **FVect** as introduced by Coecke et al. [6], and via conceptual spaces and the category **ConvexRel** as introduced by Gärdenfors [11] and Bolt et al. [3]. The former allowed us to compare the meanings of sentences between languages and calculate similarity scores, and the latter allowed us to focus more on the meaning behind nouns and calculate distances between concepts.

These results are really only the beginning of what can be achieved using the DisCoCat model of meaning, however as the old Irish proverb goes:

> "Tús maith leath na hoibre."
> - *A good start is half the work.*

Appendices

A. Corpus for Vector Space Model of Meaning (English)

The following is a summary of *Star Wars: Epsiode III - Revenge of the Sith*, obtained from Wikipedia [29] and edited by the author. Note that we are making some assumptions in using this corpus. The author is assuming the model of meaning can understand third-person sentences as if they were first-person sentences; i.e. "she is pregnant" is understood to be "Padmé is pregnant". We are also assuming the model can understand sentences with conjunction; e.g. "Anakin and Obi-Wan are known for their bravery" is "Anakin is known for his bravery" and "Obi-Wan is known for his bravery". We assume the model can understand the use of the present participle, i.e. "After infiltrating General Grevious' flagship" is understood to be "After Anakin and Obi-Wan infiltrate General Grevious' flagship". Finally we also assume the corpus has been lemmatised for *Sections 3 & 4*.

It is true that some of these assumptions might be difficult to work into the vector space model of meaning, however the author feels the use of this corpus gives good examples in *Sections 3 & 4* while still being interesting for humans to parse. *Corpora A.1 & B.1* can be rewritten such that the above assumptions are no longer necessary, however the story becomes tedious to read.

Corpus A.1.

Palpatine is a mastermind who turns Anakin to the dark side of the Force.

The galaxy is in a state of civil war. Jedi Knights Obi-Wan Kenobi and Anakin Skywalker lead a mission to rescue the kidnapped Supreme Chancellor Palpatine from the cyborg General Grievous, who is a Seperatist commander. Anakin and Obi-Wan are known for their bravery and skill. After infiltrating General Grievous's flagship, the Jedi duel Dooku, whom Anakin eventually executes at Palpatine's urging. General Grievous escapes the battle-torn cruiser, in which the Jedi crash-land on Coruscant. There Anakin reunites with his beautiful wife, Padmé Amidala, who reveals that she is pregnant. While initially excited, the prophetic visions that Anakin has cause him to worry. He believes Padmé will die in childbirth.

Palpatine appoints Anakin to the Jedi Council as his representative. The Jedi do not trust Palpatine as they believe he is too powerful. The Council orders Anakin to spy on Palpatine, his friend. Anakin begins to turn away from the Jedi because of this. Meanwhile the Jedi are searching for a Sith Lord. A Sith Lord is an evil person who uses the dark side of the Force, and the Jedi try prevent anyone from turning to the dark side of the Force and to evil. Palpatine tempts Anakin with secret knowledge of the dark side of the Force, including the power to save his loved ones from dying. Meanwhile, Obi-Wan travels to confront General Grievous. The Jedi and General Grievous duel and Obi-Wan fights bravely. Obi-Wan wins his duel against General Grievous. The Jedi Yoda travels to Kashyyyk to defend the planet from invasion. The mastermind Palpatine eventually reveals that he is a powerful Sith Lord to Anakin. Palpatine claims only he has the knowledge to save Padmé from death. Anakin turns away from Palpatine and reports Palpatine's evil to the Jedi Mace Windu. Mace Windu then bravely confronts Palpatine, severely disfiguring him in the process. Fearing that he will lose Padmé, Anakin intervenes. Anakin is a powerful Jedi and he severs Mace Windu's hand. This distraction allows Palpatine to throw Mace Windu out of a window to his death. Anakin turns himself to the dark side of the Force and to Palpatine, who dubs him Darth Vader. Palpatine issues Order 66 for the clone troopers to kill the remaining Jedi, then dispatches Anakin with a band of clones to kill everyone in the Jedi Temple.

Anakin ventures to Mustafar and massacres the remaining Separatist leaders hiding on the volcanic planet, while Palpatine addresses the Galactic Senate. He transforms the Republic into the Galactic Empire and declares himself Emperor Palpatine.

Obi-Wan and Yoda return to Coruscant and learn of Anakin's betrayal against them. Obi-Wan leaves to talk to Padmé. He tries to convince her that Anakin has turned to the dark side of the Force; that Anakin has turned to evil. A brave Padmé travels to Mustafar and implores Anakin to abandon the dark side of the Force. Anakin refuses to stop using the dark side of the Force and sees Obi-Wan hiding on Padmés ship. Anakin angrily chokes Padmé into unconsciousness. Obi-Wan duels and defeats Anakin. Obi-Wan severs both of his legs and leaves him at the bank of a lava river where he is horribly burned. Yoda duels Emperor Palpatine on Coruscant until their battle reaches a stalemate. Yoda is a powerful Jedi, but he cannot defeat the evil Emperor Palpatine. Yoda then flees with Bail Organa while Palpatine travels to Mustafar. Emperor Palpatine uses the dark side of the Force to sense Anakin is in danger.

Obi-Wan turns to Yoda to regroup. Padmé gives birth to a twin son and daughter whom she names Luke and Leia. Padmé dies of sadness shortly after. Palpatine finds a horribly burnt Anakin still alive on Mustafar. After returning to Coruscant, Anakins mutilated body is treated and covered in a black armored suit. Palpatine lies to Anakin that he killed Padmé in his rage. Palpatine is an evil mastermind and leaves Anakin feeling devastated. Palpatine has won; the dark side of the Force now flows through Anakin. Meanwhile, Obi-Wan and Yoda work to conceal the twins from the dark side of the Force, because the twins are the galaxy's only hope for freedom. Yoda exiles himself to the planet Dagobah, while Anakin and the Emperor Palpatine oversee the construction of the Death Star. Bail Organa adopts Leia and takes her to Alderaan. Obi-Wan travels with Luke to Tatooine. There Obi-Wan intends to bravely watch over Luke and his step-family until the time is right to challenge the Empire. □

B. Corpus for Vector Space Model of Meaning (Irish)

For the sake of completeness we give the full Irish corpus whose translated meaning replicates *Corpus A.1*.

Corpus B.1. Is máistirmind a casann Anakin go taobh dorcha na Fórsa é Palpatine.

Tá an réaltra i stát cogaidh shibhialta. Rinne Ridirí Jedi Obi-Wan Kenobi agus Anakin Skywalker misean chun an Seansailéir Uachtarach Palpatine a shábháil ón gCeborg Ginearál Grievous, ceannasaí Seperatist é. Aithnítear Anakin agus Obi-Wan dá a gcrógacht agus dá scileanna. Tar éis longcheannais Ginearál Grievous a ionsíothláit, troid na Jedi le Dooku, a mhoraíonn Anakin ar deireadh thiar ar mholadh Palpatine. Éalaíonn an Ginearál Grievous ón t-éadromaire caithe, ina dturlingíonn na Jedi chun talamh Coruscant. Ansin, tagann Anakin le chéile lena bhean álainn, Padmé Amidala, a léiríonn go bhfuil sí ag iompar clainne. Cé go bhfuil Anakin ar bís ar dtús, tugann a fhíseanna fáidhiúla cúis imní dó. Creideann sé go gheobhaidh Padmé bás i mbreithe clainne.

Ceapann Palpatine Anakin chuig Chomhairle na Jedi mar ionadaí. Níl muinín ag na Jedi a bheith Palpatine mar a chreideann siad go bhfuil sé ró-chumhachtach. D'ordaíonn an Chomhairle Anakin a dhéanann spiaireacht ar Palpatine, a chara. Casann Anakin as an Jedi as seo. Idir an dá linn tá na Jedi ag cuardach do Tiarna Sith. Is duine olc é Tiarna Sith a úsáideann an taobh dorcha den Fhórsa, agus déanann na Jedi iarracht a chur ar dhuine ar bith a bheith ag casadh go taobh dorcha na Fórsa agus go holc. Tacaíonn Palpatine Anakin le heolas rúnda ar thaobh dorcha na Fórsa, lena n-áirítear an chumhacht chun a mhuintir a shábháil ó bhás. Idir an dá linn, téann Obi-Wan chun dul i ngleic leis an Ginearál Grievous. Troideann an Jedi agus Ginearál Grievous agus tá Obi-Wan ag troid go crua. Buaileann Obi-Wan a chath i gcoinne Ginearál Grievous. Téann Jedi Yoda go Kashyyyk chun an phláinéid a chosaint ó ionradh. Léiríonn an máistirmind Palpatine sa deireadh gurb é Tiarna cumhachtach Sith é go Anakin. Éilíonn Palpatine ach go bhfuil eolas air amháin Padmé a shábháil ón mbás. Casann Anakin i gcoinne Palpatine agus tuairiscíonn sé olc Palpatine chuig an Jedi Mace Windu. Tabhair Mace Windu aghaidh cróga ar Palpatine, agus é a dhíshealbhú go mór sa phróiseas. Ag eagla go gcaillfidh sé Padmé, idirghabhann Anakin. Is Jedi cumhachtach é Anakin agus sealaíonn sé lámh Mace Windu. Tugann an t-imréiteach seo do Palpatine Mace Windu a chaitheamh as fuinneog go dtí a bhás. Casann Anakin féin go taobh dhorcha na Fórsa agus chuig Palpatine, a ainm Darth Vader dó. Eisíonn Palpatine Ordú 66 do na trúpaí clón chun na Jedi atá fágtha a mharú, agus ansin cuireann sé Anakin le banna cluainé chuig an Teampaill Jedi a chuir bás ar gach duine. Taistilíonn Anakin go Mustafar agus maisíonn na ceannairí Separatist atá fágtha i bhfolach ar an phláinéid volcanach, agus tugann Palpatine aitheasc don Seanad Réaltrach. Athraíonn sé an Poblacht isteach sa Impireacht Réaltrach agus dearbhaíonn sé féin an tUasal Palpatine.

Fágann Obi-Wan agus Yoda go Coruscant agus foghlaimíonn siad bradú Anakin i gcoinne iad. Fágann Obi-Wan labhairt le Padmé. Déanann sé iarracht a chur ina luí di go bhfuil Anakin tar éis casadh go taobh dorcha na Fórsa; go bhfuil Anakin tar éis casadh go holc. Taistealaíonn Padmé cróga go Mustafar agus cuireann sí ar Anakin an taobh dorcha den Fhórsa a thréigean. Diúltaíonn Anakin gan stop a bhaint as an taobh dorcha den Fhórsa agus feiceann sé Obi-Wan i bhfolach ar long Padmé. Tachtaíonn Anakin Padmé feargach go neamhfhiosach. Troideann Obi-Wan Anakin agus buaileann sé. Freastalaíonn Obi-Wan dá chuid cosa agus fágann sé é i mbruach abhainn lava ina dhóitear go mór. Troideann Yoda an t-Impire Palpatine ar Coruscant go dtí go dtarlaíonn an cath mar gheall air. Is Jedi cumhachtach é Yoda, ach ní féidir leis an olc Impire Palpatine a chosc. Téann Yoda ansin le Bail Organa agus téann Palpatine chuig Mustafar. Úsáideann an t-Impire Palpatine taobh dorcha na Fórsa le tuiscint go bhfuil Anakin i mbaol.

Casann Obi-Wan go Yoda chun athghrúthú. Tugann Padmé dá mhac agus d'iníon dúbailte a n-ainmníonn sí Luke agus Leia. Braitheann Padmé brón go gairid ina dhiaidh. Faigheann Palpatine Anakin dóite go fóill fós beo ar Mustafar. Tar éis dó dul ar ais chuig Coruscant, déileálfar le comhlacht máinliachta Anakin agus clúdaítear é in oireann armúrtha dubh. Bíonn Palpatine ag Anakin go maraíodh Padmé ina chlog. Is máistirmind olc é Palpatine agus fágann mothú Anakin ar a chéile. Bhuaigh Palpatine; tá taobh dorcha na Fórsa anois ag Anakin. Idir an dá linn, oibríonn Obi-Wan agus Yoda chun na cúpla a cheilt ó thaobh dorcha na Fórsa, toisc gurb é na cúpla is dóchas ach amháin le haghaidh saoirse. Téann Yoda féin leis an bplainéad Dagobah, agus maoiríonn Anakin agus an t-Impire Palpatine an déantús an Death Star. Uchtaíonn Bail Organa Leia agus tógann sí í chuig Alderaan. Taistealaíonn Obi-Wan le Luke go Tatooine. Tá sé i gceist ag Obi-Wan féachaint go láidir ar Luke agus ar a theaghlach go dtí go mbeidh an t-am ceart dúshlán a thabhairt don Impireacht. □

References

[1] Bargelli, D. & Lambek, J. *An Algebraic Approach To French Sentence Structure*, International Conference on Logical Aspects of Computational Linguistics, pp. 62-78. Springer, Berlin, Heidelberg, (2001).

[2] Bargelli, D. & Lambek, J. *An Algebraic Approach to Arabic Sentence Structure*, Linguistic Analysis **31**, pp. 301-315, (2003).

[3] Bolt, J., Coecke, B., Genovese, F., Lewis, M., Marsden, D. and Piedeleu, R. *Interacting Conceptual Spaces I: Grammatical Composition of Concepts*, arXiv:1703.08314v2, (2017).

[4] Caraballo, S. *Automatic construction of a hypernym-labeled noun hierarchy from text*, Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, (1999).

[5] Casadio, C. and Lambek, J, *A Computational Algebraic Approach to Latin grammar*, Research on Language and Computation **3**(1) pp. 45-60, (2005).

[6] Coecke, B., Sadrzadeh, M. and Clark, S. *Mathematical foundations for a compositional distributional model of meaning*, arXiv:1003.4394, (2010).

[7] Coecke, B., Genovese, F., Lewis, M. and Marsden, D. *Generalized Relations in Linguistics and Cognition*, Logic, Language, Information, and Computation, pp. 256-270, Springer Berlin Heidelberg, (2017).

[8] Dixon, R. M. and Aikhenvald, A. Y. *Adjective Classes: A Cross-Linguistic Typology*, Oxford University Press, (2004).

[9] Forth, J., Geraint A. W. and McLean, A. *Unifying conceptual spaces: Concept formation in musical creative systems*, Minds and Machines **20**(4), pp. 503-532, (2010).

[10] Gärdenfors, P. *Conceptual spaces as a framework for knowledge representation*, Mind and Matter **2**(2), pp. 9-27, (2004).

[11] Gärdenfors, P. *Conceptual Spaces: The Geometry of Thought*, The MIT Press, (2004).

[12] Gärdenfors, P. *The Geometry of Meaning: Semantics based on Conceptual Spaces*, The MIT Press, (2014).

[13] Grefenstette, E. and Sadrzadeh, M. *Experimental support for a categorical compositional distributional model of meaning* Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 13941404, (2011). arXiv:1106.4058.

[14] Grefenstette, E., Sadrzadeh, M., Clark, S. Coecke, B. and Pulman, P. *Concrete Sentence Spaces for Compositional Distributional Models of Meaning*, Computing meaning, pp. 71-86, Springer, Dordrecht, (2014).

[15] Grefenstette, E. and Sadrzadeh, M. *Concrete Models and Empirical Evaluations for the Categorical Compositional Distributional Model of Meaning*, Computational Linguistics, **41**(1), pp. 71-118, (2015).

[16] Gruenstein, A. *Learning Hypernyms from Corpora*, available at `citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.4605`, (2001).

[17] Hearst, M. *Automatic Acquisition of Hyponyms from Large Text Corpora*, Proceedings of the Fourteenth International Conference on Computational Linguistics, (1992).

[18] Houben, J. *A geometrically inspired category as a meaning space for natural language*, MSc thesis, available at `cs.ox.ac.uk/people/bob.coecke/Theses.html`, (2017).

[19] ItalWordNet; Italian WordNet, available at `www.ilc.cnr.it/iwndb/iwndb_php/`, accessed 25/03/2018.

[20] Lambek, J. *From Word to Sentence: a Computational Algebraic Approach to Grammar*, Polimetrica, (2008).

[21] Lambek, L. and Preller, A. *An Algebraic Approach to the German Sentence*, Linguistic Analysis, **31**(3/4), pp. 17-37, (2004).

[22] Mamlouk, A. M. *Quantifying olfactory perception*, MSc Thesis, University of Lubeck, Germany (2002).

[23] Marcus, M. P., Marcinkiewicz, M. A. and Santorini, B. *Building a large annotated corpus of English: The Penn Treebank*, Computational linguistics, **19**(2), pp. 313-330, (1993).

[24] Marsden, D. and Genovese, F. *Custom Hypergraph Categories via Generalized Relations*, arXiv:1703.01204, (2017).

[25] Princeton University, Computer Science Department. *Figure 4*, `www.cs.princeton.edu/courses/archive/spring07/cos226/assignments/wordnet.html`, accessed 20/03/2018.

[26] Riloff, E. and Shepherd, J. *A Corpus-Based Approach for Building Semantic Lexicons*, Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2), (1997). arXiv:cmp-lg/9706013.

[27] Sadrzadeh, M., Clark, S. and Coecke, B. *The Frobenius anatomy of word meanings I: subject and object relative pronouns*, Journal of Logic and Computation, **23**(6), pp. 1293-1317, (2013). arXiv:1404.5278v1.

[28] Scannell, K. *Líonra Séimeantach na Gaeilge*, available at `borel.slu.edu/lsg`, (2006).

[29] Wikipedia, *Star Wars: Episode III - Revenge of the Sith*, available at `bit.ly/2a7ZSYJ`, accessed 20/03/2018.

[30] WoNeF; WordNet *du Français*, available at `wonef.fr/`, accessed 25/03/2018.

[31] WordNet, *Princeton University "About WordNet."*, available at `wordnet.princeton.edu`, (2010).