The Syntax of Idioms

NWO/FWO-funded research project

I Research team

- Prof. dr. Norbert Corver (UIL-OTS, Utrecht University)
- Prof. dr. Jeroen van Craenenbroeck (CRISSP, KU Leuven HUBrussel)
- Dr. Tanja Temmerman (CRISSP, KU Leuven HUBrussel)
- Dr. Loes Koring (UIL-OTS, Utrecht University)

II Summary

Every language contains idiomatic expressions. In spite of their widespread occurrence, their linguistic properties have so far not been deeply and systematically investigated. The aim of this project is to increase our understanding of 'the language of idioms' by focusing on their behavior in syntax. Therefore, the main research question is: What is the syntax of idioms? This question can be naturally divided into two sub-questions, which correspond to two sub-projects: (i) What is the internal syntax of idioms, i.e. what characterizes their internal organization and makeup?; (ii) what is the external syntax of idioms, i.e. how does material contained within the idiom interact with material that is not part of the idiom? Subproject 1 aims to show that idioms, in spite of their superficially 'special' appearance, are built up by the very same syntactic and morphological structure building mechanisms that are responsible for non-idiomatic expressions. Subproject 2 aims to show that idiomatic expressions are essentially opaque domains, but that dependencies with idiom-external material is possible under certain narrowly defined circumstances. The empirical basis for our investigation into the syntax of idioms will consist of idioms from non-standard varieties of Dutch. On the basis of dialect grammars and dictionaries, and also on the basis of additional fieldwork, an electronic database will be developed which will serve as a tool for our systematic investigation into the syntax of idioms.

III Summary in key words

idioms, syntax, compositionality, opacity, Dutch dialects

IV Research proposal: The syntax of idioms

IV.1 Theoretical background and main research question

Every language contains so-called fixed expressions, i.e. combinations of words or morphemes the meaning of which cannot be deduced from those of its component parts. More formally, these expressions do not adhere to the Fregean principle of compositionality. In (1), for example, the idiomatic meaning of the entire sentence does not follow from the meaning of the words making up that sentence.

(1) The shit hit the fan.

Expressions of this type come in many shapes and sizes (ranging from entire sentences to individual words), and accordingly, they go by many different names in the linguistic literature. For expository purposes, we use the term 'idiom' in this project proposal as a general cover term.

One aspect of idioms that has received ample attention in the literature concerns the question of whether—and if so, how—they are stored in our mental lexicon (Jackendoff 1995, Marantz 1997, Everaert 2010). Given that the meaning of idiomatic expressions has to be learned much like a language learner has to learn and store the meaning of an individual word like *book*, there is a sense in which the two should be regulated by the same grammatical module, i.e. the lexicon. What is very often missing in this line of research, however, is an in-depth consideration of the *syntax* of idioms. It is precisely this void that the present project intends to fill. Its main research question can be formulated as follows:

(2) **Main research question**
What is the syntax of idioms?

The **theoretical innovation** of this project is that it investigates idioms as prime examples of the **interaction between compositionality and opacity**. Consider the following example:

(3) John kicked the bucket.

On the one hand, the idiom *kicked the bucket* is a fully regular English verb phrase (VP). In particular, the verb *kicked* receives the same past tense ending as it would in a non-idiomatic context, the order of verb and object (the former preceding the latter) is perfectly in line with the grammar of English, and the combination of a transitive verb and a definite singular object leads to an aspectual interpretation of telicity, again completely as expected (Nunberg 1978, McGinnis 2002, Everaert 2010). On the other hand, however, this completely regular VP also shows clear signs of irregularity/opacity. As soon as one of its component parts is modified, deleted or substituted, the idiomatic meaning disappears. This is illustrated in (4).

(4) a. John kicked buckets. (no idiomatic meaning)
b. John kicked his bucket. (no idiomatic meaning)
c. John kicked the can. (no idiomatic meaning)

This dual nature, we will argue, is what characterizes idioms. More specifically, we want to pursue the idea that an idiom is regular/compositional as far as its *internal* syntax is concerned, but irregular/opaque as far as its *external* syntax is concerned. Accordingly, we formulate our main research hypothesis as in (5).

(5) **Main research hypothesis**
Idioms have a compositional syntax internally, but an opaque syntax externally.

The **empirical innovation** of this project concerns the type of data it will bring to bear on the research question in (2). Existing discussions of idioms typically focus on English, and within that language on a relatively small set of 'poster children' (the examples in (1) and (3) are among that select group, along with phrasal verbs, cf. Chomsky 1981 among many others). Needless to say, it is hard to find reliable syntactic generalizations in such a small and relatively unvaried data set. This project wants to tap into an existing, but hitherto completely unexplored source of data, i.e. **non-standard varieties of Dutch**. The reasoning is simple: many of the dialects of Dutch have been described in dialect grammars (cf. http://www.meertens.knaw.nl/projecten/sand/bibliografie/sandbiblioframe.html for an overview of the existing sources), and while these works differ greatly in length, depth and quality, there is one constant: virtually all of them provide an inventory of idioms and fixed expressions typical for that dialect. In this project we want to explore this extensive data collection and use it as a basis for our theoretical discussion of idioms.

IV.2 Regularity vs. opacity: Two subprojects

As the hypothesis formulated in (5) makes clear, this project on the syntax of idioms naturally splits into two subprojects. The first one will focus on the internal syntax of idioms, the second one on their external syntax. We introduce each subproject in turn.

IV.2.1 Subproject #1: The internal syntax of idioms

The fixed choice of lexical expressions in idioms (see (4)) makes it non-trivial to explore (the regularity of) their internal syntax. Nevertheless, there are three areas that make such an exploration possible: constituency, compositionality, and morphological idioms. We discuss each in turn.

If idioms are built up internally by the very same syntactic mechanisms that create non- idiomatic expressions, then the null assumption is that constituency works similarly as well. More in particular, if an idiom is built up by a contiguous series of syntactic operations, it should form a constituent. In this respect, the first subproject will pay great attention to idioms such as the ones illustrated in (6) and (7) ((6) is taken from Ternat Dutch, cf. Van der Cruys 2008:18).

(6) Jef eid in mèn ruipe geskeetn.
Jef has in my turnips shit
'Jef has wronged me.'

(7) Dat weet mijn [neus/kleine teen/grootje/kleine broertje/hond] ook.
that knows my [nose/little toe/grandma/little brother/dog] too
'That's trivial.'

(8) Blankenberge Dutch: *kleine teen* 'little toe', *hond* 'dog'
North-Brabantish: *grootje* 'grandma', *kleine broertje* 'little brother'

What is crucial in (6) is that the possessor of *ruipe* 'turnips' is not part of the idiom; it can vary freely and is not subject to strict lexical restrictions of the type shown in (4). Put differently, the idiom seems to be *in...ruipe skauitn* 'in...turnips shit', but this is precisely the type of discontinuous non-constituent that regular(ly compositional) syntactic principles cannot create. (7) shows a different type of discontinuity. Speakers typically allow a variety of NPs as the subject of the otherwise completely frozen sentence (e.g. the topicalization of *dat* 'that' is obligatory). Moreover, there is regional variation in the NPs which can be selected to realize this subject, as shown in (8). As such, examples like these present a potential source of counterevidence against the central hypothesis pursued in this subproject. The analysis that will be offered will examine to what extent a syntactic skeleton—e.g. the structural possessor position inside the prepositional phrase in (6)—can be part of an idiom.

The second testing ground for the internal syntax of idioms concerns the presence of compositionality. At first sight, this looks like a contradiction in terms, because we have defined idioms as expressions with a non-compositional meaning. However, they sometimes display what one could call 'idiomatic compositionality', i.e. the idiomatic meaning can be decomposed into other idiomatic meanings (Nunberg et al. 1994). An example is given in (9).

(9) een appeltje voor de dorst
an apple for the thirst
'something that is saved for a rainy day'

The idiomatic expression in (9) is clearly decomposable into *een appeltje* 'an apple', which represents that which is saved, and *voor de dorst* 'for the thirst', which refers to the difficult times ahead. Importantly, though, neither *een appeltje* nor *voor de dorst* is an idiom in and of itself:

(10) # Jan heeft een appeltje gespaard. (no idiomatic reading)
Jan has an apple saved

(11) # Jan heeft wat geld op de bank staan voor de dorst. (no idiomatic reading)
Jan has some money on the bank stand for the thirst

This means that the example in (9) contains a single idiom, and more importantly, that this idiom is decomposable. Given that compositionality is the hallmark of regular syntactic structure building mechanisms, examples such as these provide evidence in favor of the hypothesis pursued in this subproject.

The third test to be applied in this subproject concerns what we will call morphological idioms, i.e. idiomatic expressions that consist of a single word. In this domain a lot of data is waiting to be uncovered. Examples are given in (12) and (13).

(12) afgezaagd
off.sawn
'hackneyed'

(13) boterbriefje
butter.letter.DIMINUTIVE
'marriage certificate'

The subproject will compare syntactic restrictions in this domain with restrictions on phrasal idioms. More specifically, we will examine whether they observe the same locality and constituency restrictions (see Embick 2010) and to what degree they may contain inflectional and other functional material (see Borer to appear)

Summing up, the first subproject focuses on the internal syntax of idioms and aims to show that these expressions are built up by the very same syntactic structure building mechanisms that are responsible for non-idiomatic expressions. Its main research question and hypothesis can be formulated as in (14) and (15). The project aims to explore this issue by focusing in particular on constituency, idiomatic compositionality and morphological idioms.

(14) **Main research question**
What is the internal syntax of idioms?

(15) **Main research hypothesis**
The internal syntax of idioms is built up through the same regular, compositional structure building mechanisms that create non-idiomatic structures.

IV.2.2 Subproject #2: The external syntax of idioms

This subproject examines the restrictions imposed on the external syntax of idioms. 'External syntax' refers to the interaction between material that is contained within the idiom and material that is not. Three types of such interaction will be investigated: binding, movement, and affixation/selection.

Coreference between nominal expressions is closely regulated in natural language (Chomsky 1981). A central notion in that debate is 'binding domain', i.e. a particular locality domain across the boundaries of which certain binding relations do not hold. What this subproject will investigate is to what extent and in what way idioms are binding domains. The data in (16) e.g. show that an idiom can prevent a proper name inside of it from being illegitimately bound by a higher nominal expression, i.e. the idiom prevents the proper name from incurring a Principle-C-violation.

(16) a. Did Adam$_i$ have an Adam$_i$'s apple?
b. * Did Adam$_i$ eat Adam$_i$'s apple?

On the other hand, examples like (17)a,b show that in some cases binding is able to cross the idiom boundary.

(17) a. Ze$_i$ praten langs elkaar$_i$ heen.
they speak along each.other across
'They're speaking at cross-purposes.'

b. * Hij praat langs elkaar heen.
he speaks along each.other across

The idiom here is *langs elkaar heen praten*. As the ungrammaticality of (17)b shows, there is a binding relation between the idiom-internal anaphor and the non-idiomatic subject. This contrasts with (16), where the idiom boundary 'blocked' binding. What these examples show is that binding sometimes can and sometimes cannot cross an idiom boundary. Thus, binding constitutes an ideal testing ground for gaining more insight into the type and degree of opacity that idioms display in their external syntax.

The second relation that can cross idiom boundaries in some, but not other cases is movement (see also Schenk 1995). Consider (18)a,b.

(18) a. Headway was made by Bill.
b. * The bucket was kicked by Bill.

These examples contain a VP-idiom (*to make headway; to kick the bucket*). Only in (18)a, however, can the direct object be promoted to subject position in a passive construction. This movement crosses the idiom boundary; as shown below, the subject position is not part of the idiom in either case.

(19) Bill/a boy/everyone/we made headway.
(20) Bill/a boy/everyone/we kicked the bucket.

Once again, the data show that idioms sometimes are and sometimes aren't transparent with respect to their external syntax. A central aim will be to delineate the precise boundaries of this phenomenon.

The third aspect of the external syntax of idioms concerns affixation/selection. It turns out that many selectional relationships—morphological and syntactic—can cross idiom boundaries, and thus reveal another breach of the opaque nature of idiomatic expressions.
Consider (1) again. At first sight, the entire sentence is part of the idiom, but this is not entirely accurate:

(21) a. The shit hits the fan.
b. The shit will hit the fan.
c. The shit has hit the fan.

The temporal and aspectual specification can vary freely and so is not part of the idiom. At first glance, this makes these examples resemble the discontinuous idioms from subproject #1, but if one assumes that the subject starts out in a position below tense and aspect (Koopman & Sportiche 1991), the constituency of *the shit hit the fan* can be retained. Importantly, however, the data in (21) show that an element outside the idiom (the tense/aspect-node) can enter into a selectional relationship with an element inside the idiom (the verb). In some cases ((1), (21)a) this results in the verb surfacing with the appropriate agreement affix, while in others ((21)bc) it surfaces as an infinitive or past participle because of the selecting auxiliary. Thus, selection represents another case where idioms seem to be transparent rather than opaque.

Besides a description of the external syntax of idioms, this project also aims to provide a theoretical account for why particular types of dependencies are blocked and others are not. The starting point for this will be the generative literature on opacity, which currently recognizes three types of opacity domains: phases (Chomsky 2000, 2001), workspaces (Uriagereka 1999, Zwart 2009, De Belder & Van Craenenbroeck 2011), and ellipsis domains (Baltin 2010, Aelbrecht 2010). Interestingly, these domains differ in their degree of opacity, i.e. in the type and number of dependencies that are blocked across the domain boundaries. This subproject will investigate which of these domains provides the best fit for idioms and will propose a theoretical analysis for the external syntax of idioms accordingly. The main research question and hypothesis are the following:

(22) **Main research question**
What is the external syntax of idioms?

(23) **Main research hypothesis**
In terms of their external syntax, idioms represent opacity domains. Their—sometimes varying—degree of opacity can be reduced to one (or more) of the known opacity domains (phases, workspaces, ellipsis domains).

IV.3 Methodology

As pointed out in section one, the main empirical innovation of this project is that it will investigate idioms from the point of view of non-standard varieties of Dutch. A large percentage of these dialects has been extensively described in dialect grammars and dictionaries, and while only few of these works devote any attention to syntax, idioms, fixed expressions and other forms of non-compositional word and sentence formation do get described in detail. This rich data source has hitherto not been explored by theoretical linguists, so a first step in the methodology of both subprojects will be to go through all these texts and create a database in which the idioms can be classified and tagged according to criteria such as categorial type (e.g. compound noun, noun phrase, verb phrase, sentence, etc.), constituency (discontinous or not), idiomatic compositionality, opacity, etc.

This database—which will be made publicly available—will not only open up these data for (further research by) the scientific community, it also forms the basis for the second methodological phase in both subprojects: the fieldwork. Based on the material in the database, the researchers will identify lacunae in the existing dialectal data set, and will construct questionnaires to fill those gaps. Those questionnaires will form the basis for oral and written interviews in the Netherlands (carried out by the Utrecht researcher) and Flanders (carried out by the Brussels researcher).

The data thus collected will serve as input for the third phase of the project: the theoretical analysis of the internal and external syntax of idioms. Starting out from the hypotheses in (15) and (23), the researchers approach the data from a theoretical perspective. The framework adopted in this phase will be Chomsky's (1995, 2000, 2001) Minimalist Program.

IV.4 Results

As will be clear from the preceding discussion, the combination of the two subprojects will yield both descriptive and theoretical results. On the one hand, the project will open up and make accessible to the scientific community a large set of language data that has so far remained unexplored. The data will be classified, tagged and collected in an electronic database which will be made available at the end of the project. This database will constitute the first ever syntactically annotated corpus of idiomatic expressions.

On the other hand, the theoretical analyses carried out by the researchers in this project will yield new insights into the basic structure building mechanisms of syntax and morphology (subproject 1) and the different types of opacity domains and concomitant varying degrees of opacity (subproject 2). The results of these studies will be published in a series of articles in international and national peer-reviewed journals and will be presented at international(ly acclaimed) linguistics conferences. Moreover, the two researchers will organize a workshop on idioms at the end of the project.