# *A Syntactically Annotated Idiom Database (SAID) v.1*

Koenraad Kuiper
Heather McCann
Heidi Quinn
Therese Aitchison: *Department of Linguistics, University of Canterbury*
and
Kees van der Veer: *Max Planck Institut für Psycholinguistic, Nijmegen.*

For e-mail contact with the authors: *<kon.kuiper@canterbury.ac.nz>*

All uses of SAID are to be bibliographically acknowledged.

## What's in SAID and how to use it.

### *1. Why?*

This data set was constructed with a number of ends in view. The chief of these was to provide data for investigating the structural configurations in which English idioms are typically found. The assumption was that, since idioms are phrasal lexical items (PLIs), they will therefore have structural properties which are idiosyncratic.

A number of these can be noted.

a. Bound words exist in a number of PLIs.

e.g.

*take umbrage*

*take cognisance of*

*have an inkling*

What are the properties of these words and how frequent are they in PLIs?

b. There may be locality constraints on the syntactic configuration of PLIs.

Are all lexicalized constituents within the maximal projection domain of the head of the PLI?

Do all lexical heads of a PLI form a lexical selection chain within the M domain of the head of the PLI? (Van Gestel 1995, O'Grady 1998, Kuiper and Everaert 2000).

c. PLIs can have slots, unfilled positions (Williams 1994)

Some are empty argument positions.

e.g. *take NP to task*

Some are not

e.g. *get NP's goat.*

Some slots have selectional restrictions.

e.g. *blow hot and cold*

(Only human subjects can blow hot and cold.)

Some have co-indexing restrictions.

e.g. *get NP's goat*

(The NP in the genitive cannot be co-referential with the subject of the PLI.)

d. Some PLI’s have constituents which may or may not be used but are known to be are part of PLI.

e.g. *Good riddance (to bad rubbish).*

e. Some PLIs have options that are a smaller set than the syntax and semantics of the PLI would generally permit.

e.g. *in a bad/foul mood/temper*

e.g. *in a good mood/temper*

but not e.g. #*in a pleasant mood*

(We use a # to indicate a phrase which is not lexicalised.)

f. Some PLIs allow the optional insertion of free modifiers (adjuncts) (Abeillé 1995: 19).

e.g. *get annoyed, get very annoyed, get slightly annoyed.*

e.g. *Get lost! #Get very lost!*

g. Some PLIs have degrees of freedom under movement with a range from frozen to free, e.g. passivisation of:

e.g. *take care of NP*

e.g. *poke borax at NP*

e.g. *kick the bucket*

(Abeillé 1995, Chafe 1968, Nunberg et al. 1994)

h. Some PLIs are restricted collocations where more than one option for selection exists in the grammar but only one is conventionally selected (Howarth 1996).

e.g. *get on the bus, #get in the bus*

e.g. *to the best of one’s abilities, #at the best of one’s abilities*

Further questions arise such as:

How much adjunction is there in PLIs?

Can any functional projection be potentially lexicalized?

Can slots occur internal to a PLI?

How are PLIs entered into the computational system in a minimalist model of syntax?□□

Data are needed to answer these and many other questions relating to the structural properties of phrasal lexical items. PLIs are also used as data in arguing for particular theoretical positions (Everaert and Kuiper 1996). Such data would be more useful if it came with syntactic annotation.

## *2. What?*

The data was originally drawn from four dictionaries of English idioms: Cowie, Mackin 1975, Cowie, Mackin and McCaig 1983, Long 1979, and Courteney 1983. Only citation forms, suitably adapted for our purposes, were used. The citation files were amalgamated. (See SAID1.txt.) The rationale for the selection was that these are among the biggest and most comprehensive listings of English idioms.

An assumption was made that many of the structural types would be represented.

No assumption was made that the selected items were a statistically significant subset of the total phrasal lexicon of English. Estimates of the size of the phrasal lexicon of an average native speaker of English range from about the same order of magnitude as the single word lexicon (Jackendoff 1995: 137) to an order of magnitude larger (Mel'çuk 1995: 169). Both these are guestimates. Given that the current guestimates of the size of the single word vocabulary of a native speaker are conservatively between 20,000 (Goulden, Nation and Read 1990) and 60,000 words (Fromkin 2000: 8), claims as to whether our sample is representative in some way would be premature.

There are 13,467 PLIs in the SAID1.txt file

## *3. Who and how?*

The analysis was conducted by Heather McCann and Koenraad Kuiper, the checking by Heidi Quinn and Therese Aitchison. Each went over the analysis of the other in the pair double-checking to attempt to gain consistency. Computational checking of bracketing was done by Kees van der Veer who also did the conversion to PROLOG and the various other formats.

The analysis was manual for the following reasons. First, when the analysis began (quite some years ago), machine parsers were not able to provide sufficient detail. Second, manual annotation raised questions about the best analysis which were heuristically challenging. Third, the time period taken for the analysis allowed a number of people to work on the project both with analysis and checking and this has led to a perhaps more considered analysis than what might have been done with faster machine parsing.

There are consequences. These data are likely to be not without error. They have been through analysis twice and checked twice. There are no bracketing errors since bracketing symmetry has been checked computationally. But it is likely that there remain errors of commission and omission. We apologize for these. In our defence we would say that, if the analysis had been done computationally, then all computer analysis errors would have been systematic and thus probably created more problems for the user than our odd, casual and unsystematic errors.

## *4. The analysis*

The analytic system we used was initially drawn from a pre-barriers generative framework. The following were notable analytic decisions:

### a. Projections and categories in the verb complex

Following Chomsky (1981), we are assuming that the basic structure of a sentence (S) is NP - AUX - VP (1).

(1) [S[NP][AUX[Vhave]][VP[Vgot][NP[DETthe][Nidea]]]]

The VP is headed by the lexical verb, and contains no further verbal elements. Any modal and non-modal auxiliaries are constituents of AUX, as is the infinitive marker *to*. As can be seen from (1), non-modal auxiliaries are assigned the category V.

Modal auxiliaries have the category MOD (2).

(2) [AP[Aas][S’[S[NP[Nchance]]**[AUX[MODwill]]**[VP[Vhave][NP[PRONit]]]]]]

The infinitive marker *to* is bracketed like an auxiliary verb, but does not have its own category label (3).

(3) [AP[Aneedless][S'[S[PRO]**[AUX[to]]**[VP[Vsay]]]]]

As (4) illustrates, AUX may contain more than one verbal element.

(4) [S[NP]**[AUX[MODmust][Vbe]]**[VP[Vseen]][S'[S[PRO]**[AUX[to][Vbe]]**[VP
[Vbelieved]]]]]

When the lexical verb *be* is inverted in WH-questions, this is treated like an auxiliary verb, as in

(5) [S'[COMP[NP[PRONwhat]]][S**[AUX[Vbe]]**[NP][VP[PP[Pin][NP[Naid]]
[PP[Pof]]]]]]

Other constituents typically found in AUX are the negative marker *not* (6), and any VP-external APs (7)-(8).

(6) [S[NP]**[AUX[MODwill][NEGnot]]**[VP[Veat][NP[PRONyou]]]]

(7) [S[NP]**[AUX[MODcan/MODcould][AP[Aalways]]]**[VP[Vdo][NP]]]

(8) [S[NP]**[AUX[MODcan/MODcould][AP[Ahardly]][Vbe]]**[VP[Vdescribed]
[PP[Pas][NP]]]]

### b. Non-finite clauses and PRO

PRO appears in the subject position of non-finite clauses that lack an overt subject and have the infinitive marker *to* in AUX. Following Chomsky (1981, 1986) we are assuming that PRO must be ungoverned. Any clause with a PRO subject is therefore presented as projecting up to S' level (9).

(9) dressed [S'[S[PRO][AUX[to]][VPkill]]]

### c. Gerund and participle constructions

Since there is considerable debate about the categorial and structural status of different gerund and participle constructions (cf. Abney 1987, Cowper 1993, Kratzer 1996), we have decided to represent gerunds as deficient clauses that project only up to S level and are therefore unable to take PRO subjects. This means that gerunds will occur either without any subject at all (10), or with overt subject NP (11). Similar constructions involving the past/passive participle of the verb are given an analogous analysis (12).

(10) keen on [S[VPing]]

(11) not see [S[NP[PRONit/that]][VP[Vhappening]]]]]

(12) have got [S[NPsb][VP[Vfooled]]]]]

### d. Small clauses

Where an NP is followed by a non-verbal predicative phrase (13)-(15), or by a VP headed by a bare infinitive (16), we have analysed the whole constituent as a small clause. To ensure easy identification and compatibility with different existing approaches to small clauses (Aarts 1992, Bowers 1993, Chomsky 1993, Cardinaletti & Guasti 1995), all small clauses identified in the dictionary have the category label SC and a basic NP - XP structure.

(13) [SC[NP[Nback]][PP[Pto][NP[Nfront]]]]

**PP predicate**

(14) [SC[NP[Nall]][AP[AP[Avery]][Afine/Awell]([PP[Pfor][NPsb]])]]

**AP predicate**

(15) [**[SC[NP[Nall]][NP[DETthe][Nsame]]]**[PP[Pto][NP]]]

**NP predicate**

(16) [S[NP][AUX[MODcan't]][VP[Vhear]**[SC[NP[PRONoneself]]**

**[VP[Vthink]]]**]]]

### e. Possessives

We are treating the possessive marker *'s* as a lexical clitic that turns an NP into a possessive phrase (cf. Halpern 1995 for a similar approach). Such possessive phrases are assigned the category POSS (17).

(17) [AP[Alike][NP**[POSS[NP[DETthe][Ncurate]]'s]**[Negg]]]

Possessive pronouns are assumed to project both NP and POSS, without the need for a separate possessive marker (17).

(18) [AP([AP[DEGas]])[Aplain][S'[COMPas][S[NP[DETthe][Nnose][PP[Pon]

[NP**[POSS[NP[PRONyour]]]**[Nface]]]]]]]

Where the possessive form of a pronoun is indicated by *'s*, the pronoun is analysed as the head of a noun phrase taking a possessive clitic (19).

(19) [AP[Aaccording][PP[Pto][NP**[POSS[NP[PRONone]]'s]**[Nlights]]]]

### f. Conjunction

Conjunction constructions are assumed to be headed by the conjuncts rather than the conjunction (cf. Pesetsky 1982, Gazdar et al. 1985, Jackendoff 1990, Pollard & Sag 1994, Kaplan & Maxwell 1995). This means that the overall constituent inherits the category from its conjuncts (20).

(20) **[AP**[AP[Aancient]][CONJand][AP[Ahonourable]]**]**

If the conjuncts belong to different categories, the overall constituent is bracketed but does not bear a category label (21).

(21) **[**[AP[Aclothed]][CONJand][PP[Pin][NP[POSS[NP[PRONone]]'s][AP[Aright]]

[Nmind]]]]**]**

Conjunctions have the category CONJ. If no overt conjunction is present, conjuncts are separated by an empty CONJ node (22).

(22) [AP[AP[Acool]]**[CONJ]**[AP[Acalm]][CONJand][AP[Acollected]]]

It is usually phrases that are conjoined, but sometimes words are conjoined, as in (23).

(23) [VP[Vweave][PP[P[Pin][CONJand][Pout]][PP[Pof][NP]]]]

### g. Comparative and equative structures

Our analysis distinguishes two types of comparative/equative structures:

**Type 1**: Comparatives and equatives where the degree adverb (*more, as, enough*) semantically modifies an A (24) or N (25). The constituent following *than* or *as* is either clausal or readily interpreted as a reduced clause.

(24) *more beautiful(ly) than ever*

*as bad as ever*

*old enough to be X*

(25) *more trouble than be worth*

Our syntactic analysis of Type 1 comparatives and equatives is designed to reflect their semantic properties. Thus the initial degree adverb is treated as a modifying AP, and *than* and the second *as* in equatives are analysed as complementizers introducing an embedded clause (26)-(28).

(26) [AP**[AP[DEGmore]]**[Abeautiful][S’**[COMPthan]**[S[AP[Aever]]]]]

[NP**[AP[DEGmore]]**[Ntrouble][S’**[COMPthan]**[S[NP][VP[Vbe]

[AP[Aworth]]]]]]

(27) [AP**[AP[DEGas]]**[Afar][S’**[COMPas]**[S[NP][AUX[MODcan]][VP[Vsee]]]]]

(28) [AP[Aold]**[AP[DEGenough]]**[S’[S[PRO][AUX[to]][VP[Vbe]

[NP[POSS[NP[PRONone]]'s][Nfather]]]]]]

Support for this analysis comes from the optionality of the initial *as* in many equative structures (29), and from the occurrence of morphological as well as periphrastic comparatives in Type 1 comparative constructions (30)-(31).

(29) [AP**([AP[DEGas]])**[Abad][S'[COMPas][S[NP[Never]]]]]

(30) [S[NP[PREDEThalf][DETa][Nloaf]][VP[Vis]**[AP[Abetter]**[S'[COMPthan]

[S[NP[QP[Qno]][Nbread]]]]**]**]]

(31) **[AP[DEGmore]**[S'[COMPthan][S[AP[Alikely]]]]**]**

[AP**[AP[DEGmore]**[S'[COMPthan][S[NP[DETa][Nbit]]]]**]**[A]]

**Type 2**: Comparatives where the degree element (*more, better*) does not have this kind of modifying relationship with the following word, but rather seems to introduce a kind of coordinate structure. For this reason, the degree element may be followed by clauses (32) as well as phrasal constituents (33). As we might expect from a structure resembling *either…or…* and *both…and…* coordinates, the constituent following *than* is usually of the same category as the constituent following the degree element.

(32) *better to be safe than sorry*

(33) *more dead than alive*

*better late than never*

*more in sorrow than in anger*

Our analysis of Type 2 comparatives captures the special status of the degree element and the symmetry between the compared constituents by assuming that the whole structure is headed by the degree adverb, which takes the following constituent and the *than* clause as its complements (34)-(35).

(34) [AP**[DEGmore]**[AP[Adead]][S'[COMPthan][S[AP[Aalive]]]]]

[AP**[Abetter]**[AP[Alate]][S'[COMPthan][S[AP[Anever]]]]]

[AP**[DEGmore]**[PP[Pin][NP[Nsorrow]]][S'[COMPthan][S[PP[Pin]

[NP[Nanger]]]]]]

(35) [AP**[Abetter]**([S'[S[PRO][AUX[to]][VP[Vbe])[AP[Asafe]]([]])[S'[COMPthan]

[S[AP[Asorry]]]]]

**h. The category of *all***

*All* is treated as a QP when it modifies an overt N or PRON (36), as an AP when it modifies an overt A (37), and as an NP when it occurs by itself (38).

(36) a. [NP**[QP[Qall]]**[DETthe][Nfun][PP[Pof][NP[DETthe][Nfair]]]]

b. [S[NP][AUX[Vhave]][VP[Vseenetc][NP[PRONit]**[QP[Qall]]**]]]

(37) [AP**[AP[Aall]]**[Aimportant]]

(38) [SC**[NP[Nall]]**[NP[DETthe][Nsame][PP[Pto][NP]]]]

i. Ungrammatical PLIs

We have analysed the ungrammatical PLIs as best we can. For example, where these are historical throwbacks, we have tried to analyse them in line with the grammar of the period as in (39)

(39) *[S[VP[Vgather]][NP[Nye]][NP[Nrosebuds]][S'[COMPwhile][S[NP[Nye]]

[AUX[MODmay]]]]]

## *5. Conventions*

The following conventions were adopted in the analysis to give SAID2.txt which is the manual analysis file.

1. Square brackets enclose constituents.
2. Upper case notation inside the leftmost bracket provides the syntactic label for the constituent.
3. All linguistic data is reduced to lower case.
4. / is placed between alternative heads (selection sets).
5. () is placed around lexicalized optional constituents.
6. NP is used for many slots.
7. Dictionary abbreviations like *sb* and *sth* for *somebody* or *something* are also used within slots.
8. * indicates an ungrammatical PLI

## 6. What is missing?

No single bar levels are used.

No traces are indicated.

No co-indexing is noted.

## *7. File types*

In order to facilitate machine manipulation of the annotated data, the manual analysis was converted to PROLOG format. This involved expansions of those PLIs which had optional constituents so that both the case with and that without the options were made available. Alternatives were left in the PROLOG file separated by /. SAID3.txt contains the data in SAID2 form above and Prolog form below for each datum. SAID4.txt contains only the PROLOG form of the data.

We think that the various file formats provided will make it possible to convert our format to others with judiciously constructed algorithms.

Files have been left in text format with each record separated by a return. This should make it easy to import the data into any database for interrogation.

We have also enclosed Theo Vosse's program 'TreeParse' and its manual, with his blessing, for Macintosh users since it will draw tree diagrams from PROLOG input.

## *8. Users*

We hope this data set will be useful for linguists. Those working in parsing and machine translation might find the data useful for priming linguistic analysis of new data and cutting down the search space for non-compositional phrases in parsing and machine translation algorithms.

Some teachers of English as a second or foreign language may also find the structural analyses useful for grounding grammar learning in idioms which are often themselves memorable or at least worth knowing if you are a foreign language learner.

## *9. Caveat*

We have made our best effort as to the consistency and accuracy of the data analysis; however no guarantees are made or implied as to either.

## *10. Thanks.*

We are grateful to the following for grants:

The New Zealand Vice Chancellors' Committee,

The University of Canterbury.

Koenraad Kuiper is grateful to:

the University of Canterbury for periods of study leave during which some of his contribution to the project was made,

NWO who provided support during a period of leave hosted by het Onderzoeksinstituut voor Taal and Spraak (OTS) at the University of Utrecht (1995-6),

the Max Planck Institut für Psycholinguistic and Professor W.J.M. Levelt for hospitality and assistance during a three month's stay at the Institute in 1999,

the Netherlands Institute for Advanced Studies in the Humanities and Social Sciences for a Fellowship in 2002-2003.

## *11. References*

Aarts, Bas. 1992. Small clauses in English: the nonverbal types. (Topics in English linguistics, 8) Berlin: Mouton de Gruyter.

Abbeillé, Anne. 1995. The flexibility of French idioms: A representation with lexicalized tree adjoining grammar. Idioms: Structural and psychological perspectives, ed. by Martin Everaert, Eric-Jan van der Linden, André Schenk, and Rob Schroeder, 15-52. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Abney, Steven Paul. 1987. The English noun phrase in its sentential aspect. Unpublished PhD thesis. Cambridge, MA: MIT.

Borsley, Robert D. 1994. In defense of coordinate structures. Linguistic Analysis 24. 218-46.

Bowers, John. 1993. The syntax of predication. Linguistic Inquiry 24. 591-656.

Cardinaletti, Anna and Maria Teresa Guasti (eds). 1995. Small clauses. (Syntax and semantics, 28) San Diego: Academic Press.

Chafe, Wallace. 19688. Idiomaticity as an anomaly in the Chomskian paradigm. Foundations of language 4. 109-127.

Chomsky, Noam. 1981. Lectures on government and binding. Dordrecht: Foris.

Chomsky, Noam. 1986. Knowledge of language: its nature, origin, and use. (Convergence) New York: Praeger.

Chomsky, Noam. 1993. A minimalist program for linguistic theory. The view from building 20: essays in linguistics in honor of Sylvain Bromberger, ed. by Kenneth Hale and Samuel Jay Keyser, 1-52. (Current studies in linguistics, 24) Cambridge, MA: MIT Press.

Courteney, Rosemary. 1983. Longman dictionary of phrasal verbs. Harlow, Essex: Longman.

Cowie, Anthony P. and Ronald Mackin. 1975. Oxford dictionary of current idiomatic English: Verbs with prepositions and particles. Oxford: Oxford University Press.

Cowie, Anthony P. Ronald Mackin and Isobel McCaig 1983. Oxford Dictionary of Current Idiomatic English: Phrase, clause and sentence idioms. Oxford: Oxford University Press.

Cowper, Elizabeth. 1993. A non-unified treatment of *-ing*. Toronto Working Papers in Linguistics 12. 49-59.

Everaert, Martin and Koenraad Kuiper. 1996. Theory and data in idiom research. Parasession on theory and data in linguistics, ed. by Michele AuCoin, Rodolfo Celis, Lise M. Dobrin, Lisa McNair, and Kora Singer (CLS 32v2), 43-57. Chicago: Chicago Linguistics Society.

Fromkin, Victoria et al. 2000. Linguistics: An introduction to linguistic theory. Oxford: Blackwell.

Gazdar, Gerald, Ewen Klein, Geoffrey K. Pullum, and Ivan A. Sag. 1985. Generalized phrase structure grammar. Cambridge, MA: Harvard University Press.

Goulden, Robin, Paul Nation, and John Read. 1990. How large can the receptive vocabulary be? Applied Linguistics 11. 341-363.

Halpern, Aaron. 1995. On the placement and morphology of clitics. (Dissertations in linguistics) Stanford, CA: Center for the Study of Language and Information (CSLI).

Howarth, Peter A. 1996. Phraseology in English academic writing: Some implications for language learning and dictionary making. Max Niemeyer Verlag: Tübingen.

Jackendoff, Ray. 1990. On Larson’s analysis of the double object construction. Linguistic Inquiry 21. 427-456.

Jackendoff, Ray S. 1995. The boundaries of the lexicon. Idioms: structural and psychological perspectives, ed. by Martin Everaert, Eric-Jan van der Linden, André Schenk, and Rob Schroeder, 133-165. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Kaplan, Ronald M. and John T. Maxwell III. 1995. Constituent coordination in lexical-functional grammar. Formal issues in lexical-functional grammar, ed. by Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, 199-210. (CSLI lecture notes, 47) Stanford: CSLI.

Kratzer, Angelika. 1996. Severing the external argument from its verb. Phrase structure and the lexicon, ed. by Johan Rooryck and Laurie Zaring, 109-137. (Studies in natural language and linguistic theory, 33) Dordrecht: Kluwer Academic.

Kuiper, Koenraad and Martin Everaert. 2000. Constraints on the phrase structural properties of English phrasal lexical items. PASE papers in language studies: Proceedings of the 8th annual conference of the Polish Association for the study of English, ed. by Rozwadowska Bozena , 151-170. Aksel: Wroclaw.

Long, Thomas H. et al. 1979. The Longman dictionary of English idioms. Harlow, Essex: Longman.

Mel'çuk, Igor 1995. Idioms: structural and psychological perspectives, ed. by Martin Everaert, Eric-Jan van der Linden, André Schenk, and Rob Schroeder, 167-232. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Nunberg, Jeffrey, Ivan Sag, and Thomas Wasow. 1994. Idioms. Language 70. 491-538.

O'Grady, William. 1998. The syntax of idioms. Natural Language and Linguistic Theory 16. 279-312.

Pesetsky, David. 1982. Paths and categories. Unpublished PhD thesis. Cambridge, MA: MIT.

Pollard, Carl and Ivan A. Sag. 1994. Head-driven phrase structure grammar. Chicago: University of Chicago Press.

Van Gestel, Frank, 1995. En bloc insertion. Idioms: structural and psychological perspectives, ed. by Martin Everaert, Eric-Jan van der Linden, André Schenk, and Rob Schroeder, 75-94. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Williams, Edwin. 1994. Thematic structure in syntax. Cambridge, Massachusetts: MIT Press.