

ANNOTATION GUIDELINES FOR COMPOUND NOUNS IN ENGLISH, DUTCH AND AFRIKAANS

These guidelines were taken and adapted from Ó Séaghdha's PhD thesis 'On Compound Semantics' (2008). They are developed to be able to describe the semantic relation between the constituents of two-noun compounds. We have only annotated those compounds that are not in the dictionary, but of which the constituent nouns can in fact be found in the dictionary. If a compound already has a gloss, we do not have to analyse it to find its meaning, but we do need to know the meaning of each constituent to be able to find the compound meaning. This means that a lot of common, lexicalised and exocentric compounds are excluded from the annotation. These compounds will be removed from the annotation data by crosschecking the data with a dictionary before the annotation commences. Should we still encounter such compounds in our data, rule 1.4 explains what to do with them.

More details on the adaptation of these guidelines can be found in chapter 4 of Verhoeven's master dissertation 'A Computational Semantic Analysis of Noun Compounds in Dutch' (2012).

A classification scheme with paraphrasing prepositions and predicates and a decision tree are made available to aid the annotators in making their acquaintance with the annotation process and guidelines. These are to be considered tools that can aid in the apprehension of the annotation process or when struggling with the classification of a certain compound. These tools can be found on the project website¹.

1. General Guidelines

The task is to annotate each compound noun N1 N2 with regard to the semantic relation that holds between the constituent nouns N1 and N2. It is assumed that compounds are either copulative or semantically right-headed.

Rule 1.1 *The general annotation format is <RELATION,DIRECTION,RULE>.*

RELATION is one of the 10 relation labels defined in section 2 of these guidelines. DIRECTION specifies the order of the constituent nouns in the chosen relation's argument structure – in particular, direction will have the value 1 if the first noun in the compound (N1) fits in the first noun slot mentioned in the rule licensing the chosen relation, and will have value 2 if the second noun in the compound (N2) fits in the rule's first noun slot. RULE is the number of the rule licensing the relation. For example:

¹ <http://tinyurl.com/aucopro>

water fern

IN,2,2.1.3.1

This aquatic water fern is a rosette plant which has dense, fibrous roots

enemy provocation

ACTOR,1,2.1.4.1

The army said at the weekend that troops had reacted to enemy provocations and intervened to protect local citizens

In the case of *water fern* the IN relation is licensed by Rule 2.1.3.1 *N1/N2 is an object spatially located in or near N2/N1*. Mapping the compound's constituent nouns onto the rule definition, we see that the first slot (N1/N2 is. . .) is filled by N2 *fern* and hence the direction is 2. For the categories BE, REL, LEX, UNKNOWN, MISTAG and NONCOMPOUND there is no salient sense of directionality, so it need not be annotated:

cedar tree

BE,2.1.1.1

On rising ground at the western end of the churchyard of St Mary's at Morpeth in Northumberland stands, sheltered by cedar trees, a funerary monument

In practice, we will assign every compound a direction to have uniformity in the encoding. Every compound from a category that has no sense of directionality (see above) will be encoded with direction 1.

In the examples of section 2 you will find the direction of the example in brackets behind the compound.

Rule 1.2 *Each compound is presented with its sentential context and should be interpreted within that context. Knowledge of other instances of the compound type are irrelevant.*

A given compound type can have different meanings in different contexts. A *school book* is frequently a book read IN school, but it could also be a book ABOUT school. A *wood table* might be a table that IS wood (BE), but it might also be a table for chopping wood on (IN). The intended meaning of a compound is often clarified by the sentence it appears in.

Rule 1.3 *Where a compound is ambiguous and is not clarified by the sentential context, the most typical meaning of the compound is favoured.*

Compound interpretation must sometimes rely on world knowledge. In these cases, the annotator will have to rely on his or her intuition. Querying Google for the most typical meaning would be a viable option, but would take too much time in the annotation process.

The compound *school book* is not clarified by a sentence such as *This is a school book*. In this case, book read IN school is the most typical interpretation. If the compound's ambiguity arises from the polysemy of a constituent, the same consideration applies. University can refer to an institution or its physical location, but in the case of *university degree* the institutional meaning must be correct as locations cannot award degrees, and the compound is labelled ACTOR.

If the meaning of the compound is unclear, the appropriate tag is UNKNOWN.

Rule 1.4 *There are number of special cases that would normally not appear in our training data. If they should be present, they are to be treated differently than other compounds, they will all be annotated REL or LEX.*

- *When a compound is used metaphorically, it will not be considered a regular compound and it should be labelled LEX.*

For example: the compound *bird brain* is often used to refer to someone stupid, not to an actual bird's brain. Luckily, a lot of metaphorical compounds have such a typical meaning that they can be found in a dictionary and will therefore not be present in the annotation data.

- *Where a compound consisting of two common nouns is used as a proper noun, it will be discarded from our annotation. Also compounds that exist of one or more proper nouns, abbreviations or acronyms will be left out. All these special cases receive the REL tag.*

Many names, while constructed from two common nouns, do not seem to encode the same kind of semantics as non-name compounds, e.g. *Penguin Books*, *Sky Television*, *Dolphin Close*, *Coronation Street*. These names encode only a sense of non-specific association between the constituents. All compounds that are used as a proper noun will therefore be classified as REL, even those that could be classified otherwise. For example: the *Telecommunications Act*, *The Old Tea Shop*, *Castle Hill*. The task of identifying these proper noun compounds should be passed on to a named entity recognition (NER) module.

Rule 1.5 *Where there is a characteristic situation or event that characterizes the semantic relation between the constituents, it is necessary to identify which constituents of the compound are participants and which roles they play. Whether such a situation exists for a given compound, and the*

roles played by its constituents in the situation, will determine which relation labels are available.

Participants take on roles that can be described as Agent, Instrument, Object or Result:

- **Agent** The instigator of the event, the primary source of energy
- **Instrument** An intermediate entity that is used/acted on by the Agent and in turn exerts force on or changes the Object; more generally, an item which is used to facilitate the event but which is not the Object
- **Object** The entity on which a force is applied or which is changed by the event and which does not exert force on any participant other than the Result. Recipients (e.g. of money or gifts, but not outcomes) also count as Objects.
- **Result** An entity which was not present before and comes into being through the event

For example, the meaning of *cheese knife* seems to involve an event of cutting, in which cheese and knife take object and instrument roles respectively. Similarly, *taxi driver* evokes an event of driving and *gevangenisbewaker* (prison guard) evokes an event of guarding. The INST and ACTOR relations apply only where such a situation or event is present and where the compound identifies its participant(s). The application of HAVE assumes that the most salient aspect of the underlying situation is possession. It is not strictly necessary to identify the precise nature of the situation or event, only to identify the general roles played by the participants.

Some role-tagged examples: *cheese_O knife_I*, *taxi_O driver_A*, *sneezing_R powder_I*, *gevangenis_O bewaker_A*. It follows from the role descriptions that locations and topics do not count as participants – compounds encoding such roles receive IN and ABOUT labels instead of the ACTOR and INST labels reserved for participants.

The participant role types are listed in order of descending agentivity. We thus have an agentivity hierarchy Agent>Instrument>Object>Result¹. This ordering plays an important role in distinguishing ACTOR compounds from INST compounds (see Rules 2.1.4 and 2.1.5). It is not necessary to annotate this information, and it is not always necessary to identify the exact participant role of a constituent, so long as the hierarchical order of the constituents can be identified. Identifying participants is only needed to distinguish between relations (ACTOR vs INST) and directionalities (see the discussion under Rule 2.1.5.2).

¹ This agentivity hierarchy was informed by the semantic roles hierarchy in Talmy, 2000. Talmy, L. (2000). 'The semantics of causation'. In: *Toward a Cognitive Semantics, Volume 1: Concept Structuring Systems*. Cambridge, MA: MIT Press.

2. Semantic Relations

2.1 Main Relations

2.1.1 BE

Rule 2.1.1.1 *X is N1 and X is N2.*

For example:

English: *woman driver, elm tree, distillation process, human being.*

Dutch: *geluidhinder, rundsvlees, bombrief, puntkomma, gastarbeider, getuige-deskundige.*

Afrikaans: *digter-skrywer, briefbom, kommapunt, gasarbeider, vroueatleet, seunsvriend, wuifgroet.*

This rule does not admit sequences such as *deputy chairman, fellow man, chief executive* or *hoofdverantwoordelijke*, where it is not correct to state that an [N1 N2] is an N1 (a chief executive is not a chief). Such sequences are not to be considered compounds, and their modifiers are to be considered (mistagged) adjectives – see Rule 2.2.1.1.

Rule 2.1.1.2 *N2 is a form/shape taken by the substance N1.*

For example:

English: *stone obelisk, chalk circle, plastic box, steel knife.*

Dutch: *gummiband, betonsteen, staalkabel.*

Afrikaans: *plastiekrekkie, glasbak, houtstoel, ysterhek, silverring.*

This rule is not very productive in Dutch since substances are most often written as adjectives, e.g. *plastieken doos, stalen mes.*

Rule 2.1.1.3 *N2 is ascribed significant properties of N1 without the ascription of identity. The compound roughly denotes “an N2 like N1”.*

For example:

English: *father figure, angler fish, chain reaction, pie chart.*

Dutch: *hagelpatroon, rondengang, manwif, mensaap.*

Afrikaans: *mensaap, vaderfiguur, kettingreaksie.*

2.1.2 HAVE

Rule 2.1.2.1 *N1/N2 owns N2/N1 or has exclusive rights or the exclusive ability to access or to use N2/N1 or has a one-to-one possessive association with N2/N1.*

For example:

(The numbers after the examples refer to the direction of the semantic relation.)

English: *army base(1)*, *customer account(1)*, *government power(1)*.

Dutch: *straatnaam(1)*, *koningsdochter(1)*.

Afrikaans: *skoolgrond(1)*, *kantooradres(1)*, *menseregte(1)*.

The term one-to-one possessive association is intended to cover cases where it seems strange to speak of ownership, for example in the case of inanimate objects (*street name*, *planet atmosphere*).

Rule 2.1.2.2 *N1/N2 is a physical condition, a mental state or a mentally salient entity experienced by N2/N1.*

For example:

English: *polio sufferer(1)*, *cat instinct(2)*, *student problem(2)*, *union concern(2)*.

Dutch: *lepralijder(1)*, *studentenprobleem(2)*.

Afrikaans: *kankerpatiënt(1)*

Rule 2.1.2.3 *N1/N2 has the property denoted by N2/N1.*

For example:

English: *water volume(1)*, *human kindness(1)*.

Dutch: *productietijd(1)*,

Afrikaans: *stooftemperatuur(1)*,

A “property” is something that is not an entity or a substance but which an entity/substance can be described as having. *Redness*, *temperature*, *dignity*, *legibility* are all examples of properties.

Rule 2.1.2.4 *N1/N2 has N2/N1 as a part or constituent.*

For example:

English: *car door(1)*, *motor boat(2)*, *cat fur(1)*, *chicken curry(2)*, *pie ingredient(1)*, *tree sap(1)*.

Dutch: *houtweefsel(1)*, *bladzijde(1)*, *moutjenever(2)*, *hamersteel(1)*, *grafzerk(1)*, *tafelblad(1)*.

Afrikaans: *deurknop(1)*, *kasdeur(1)*, *storkop(1)*, *geweerloop(1)*, *sjokoladekoek(2)*, *melktert(2)*.

The test for the presence of a part-whole relation is whether it seems natural and accurate in the context to say *The N1/N2 has/have N2/N1* and *The N1/N2 is/are part of N2/N1*. Furthermore, substances which play a functional role in a biological organism are classed as parts: *human blood*, *tree sap*, *whale blubber*. This is the case even when the substance has been extracted, as in *olive oil*. A part is often located in its whole, but in these cases the part-whole relation is to be considered as prior to the co-location, and HAVE is preferred to IN. Complications arise with cases such as *sea chemical*, where both HAVE and IN seem acceptable. One principle that can be used tests whether the candidate part is readily separated (perceptually or physically) from the candidate whole. Chemicals in *sea water* (HAVE) are not typically separable in this way and can be viewed as parts of a whole. On the other hand, a *sea stone* or a *sea (oil) slick* are perceptually distinct and physically separable from the sea and are therefore IN.

Rule 2.1.2.5 *N1/N2 is a group/society/set/collection of entities N2/N1*

For example:

English: *stamp collection(2)*, *character set(2)*, *lecture series(2)*, *series lecture(1)*, *committee member(1)*, *infantry soldier(1)*.

Dutch: *postzegelverzameling(2)*, *schoenenhoop(2)*, *groepslid(1)*.

Afrikaans: *seëlversameling(2)*, *keramiekversameling(2)*, *studentegroep(2)*.

2.1.3 IN

In the following rules, an opposition is drawn between events/activities and objects. The class of events includes temporal entities such as times and durations. Objects are perceived as non-temporal and may be participants in an event (the term participant is used as defined under Rule 1.5). To assign the correct rule, the annotator must decide whether the located thing is an event or an object, and whether the location is temporal or spatial. Events may also sometimes be participants – in the sense of Rule 1.5 and in these cases the rules dealing with objects and participants will apply – a *nursing college* is a college where nursing is taught as a subject, but not necessarily one where the activity of nursing takes place, so Rule 2.1.3.1 applies. In contrast a *nursing home*, being a home where the event of nursing takes place, would come under Rule 2.1.3.2, analogous to *dining room*. Some nouns are polysemous and can refer to both objects (*play* as a written work, *harvest* as harvested crops) and events (*play* as performance, *harvest* as activity). The annotator must decide whether the temporal or physical aspect is primary in a given context.

Rule 2.1.3.1 *N1/N2 is an object spatially located in or near N2/N1.*

For example:

English: *forest hut(2)*, *shoe box(1)*, *side street(2)*, *top player(2)*, *crossword page(1)*, *hospital doctor(2)*, *sweet shop(1)*.

Dutch: *waterplant(2)*, *rivierleem(2)*, *ziekenhuisbed(2)*, *havenkantoor(2)*, *kerkdief(2)*.

Afrikaans: *vleismark(1)*, *hospitaalbed(2)*, *begrafenisrys(2)*.

Where the location is due to part-whole constituency or possession, HAVE is preferred (as in *car door*, *sea salt*). Source-denoting compounds such as *country boy* and *spring water* are classed as IN as the underlying relation is one of location at a (past) point in time.

Rule 2.1.3.2 *N1/N2 is an event or activity spatially located in N2/N1.*

For example:

English: *dining room(1)*, *hospital visit(2)*, *sea farming(2)*, *football stadium(1)*.

Dutch: *biljartzaal(1)*, *distributiecentrum(1)*, *tuinfeest(2)*, *zeeslag(2)*.

Afrikaans: *plaasbesoek(2)*, *wildtuintoer(2)*, *harsingontsteking(2)*.

Rule 2.1.3.3 *N1/N2 is an object temporally located in or near N2/N1, or is a participant in an event/activity located there.*

For example:

English: *night watchman(2)*, *coffee morning(1)*.

Dutch: *nachtuil(2)*, *sterrennacht(1)*, *lenteweertje(2)*, *weekblad(2)*.

Afrikaans: *dagblad(2)*, *nagapie(2)*, *maanskynaand(1)*.

Rule 2.1.3.4 *N1/N2 is an event/activity temporally located in or near N2/N1.*

For example:

English: *future event(2)*, *midnight mass(2)*.

Dutch: *avondfeest(2)*, *nachtvoorstelling(2)*, *jaarvergadering(2)*.

Afrikaans: *rugbyseizoen(1)*, *somerdiens(2)*.

2.1.4 ACTOR

The distinction between ACTOR and INST is based on sentience. Only certain classes of entities may

be actors:

1. Sentient animate lifeforms: membership of the animal kingdom (*regnum animalia*) is a sufficient condition. Bacteria and viruses are not sentient enough (flu virus is annotated INST).
2. Organisations or groups of people: for example *finance committee*, *consultancy firm*, *manufacturing company*, *council employee*. Some words referring to institutions are polysemous in that they can denote its physical aspect or its social/organisational aspect – university often denotes a physical location, but in the compounds *university degree* and *university decision* it is functioning as an organisation and count as agents (granting a degree and making a decision are actions only humans or organisations can carry out). On the other hand, in *research university* it is not clear whether we have a university that does research (agentive) or a university in which research is done (non-agentive). In such cases, the physical denotation should be considered the primary meaning of the word, and the organisational denotation is derived through metonymy – the non-agentive interpretation of these compounds is favoured unless the underlying event requires the institution to act as an agent. Such events often involve the institution acting as a legal entity. Hence *university degree* (degree awarded by a university), *school decision* (decision made by a school), *shop employee* (employee employed by a shop) are ACTOR; *research university*, *community school*, *school homework* and *sweet shop* are IN.

A compound can be labelled ACTOR only if the underlying semantic relation involves a characteristic situation or event. In the following definitions, the term participant is used in the sense of Rule 1.5.

Rule 2.1.4.1 *N1/N2 is a sentient participant in the event N2/N1.*

For example:

English: *student demonstration(1)*, *government interference(1)*, *infantry assault(1)*.

Dutch: *burgeroorlog(1)*, *arbeidsvrouw(2)*, *aanslagpleger(2)*.

Afrikaans: *werkerstaking(1)*, *vrouekonferensie(1)*.

That N2/N1 denote an event is not sufficient for this rule – it must be the characteristic event associated with the compound. Hence this rule would not apply to a *singing teacher*, as the characteristic event is teaching, not singing. Instead, Rule 2.1.4.2 would apply. As only one participant is mentioned in the current rule 2.1.4.1, there is no need to establish its degree of agentivity.

Rule 2.1.4.2 *N1/N2 is a sentient participant in an event in which N2/N1 is also a participant, and N1/N2 is more agentive than N2/N1.*

For example:

English: *honey bee*(2), *bee honey*(1), *company president*(2), *history professor*(2), *taxi driver*(2), *student nominee*(1).

Dutch: *aasdier*(2), *hartendief*(2).

Afrikaans: *spankaptein*(2), *voortrekkerleier*(2).

Relative agentivity is determined by the hierarchy given under Rule 1.5. The underlying event cannot be one of possession (*car owner* = HAVE) or location (*city inhabitant* = IN). Profession-denoting compounds often have a modifier which is a location – *street cleaner*, *school principal*, *restaurant waitress*, *school teacher*. A distinction can be drawn between those where the profession involves managing or changing the state of the location, i.e. the location is an object (*school principal*, *street cleaner* = ACTOR), and those where the profession simply involves work located there (*school teacher*, *restaurant waitress* = IN by Rule 2.1.3.1). Note that modifiers in *-ist* such as *expressionist*, *modernist*, *socialist*, *atheist* are treated as nouns, so that an *expressionist poem* is analysed as a poem such as an expressionist would characteristically write.

2.1.5 INST

The name INST(rument) is used to distinguish this category from ACTOR, though the scope of the category is far broader than traditional definitions of instrumentality. Again, the term participant is used in the sense of Rule 1.5.

Rule 2.1.5.1 *N1/N2 is a participant in an activity or event N2/N1, and N1/N2 is not an ACTOR.*

For example:

English: *skimming stone*(2), *gun attack*(1), *gas explosion*(1), *combustion engine*(2), *drug trafficking*(1), *rugby tactics*(2), *machine translation*(1).

Dutch: *smaakbederf*(1), *zaadhandel*(1), *leengoed*(2).

Afrikaans: *beesveiling*(1), *bomdril*(1), *ontstekingsknoppie*(2).

Compounds identifying the location of an event (such as *street demonstration*) should be labelled IN by Rule 2.1.3.2 or 2.1.3.4, and compounds identifying the focus of or general motivation for a human activity or mental process (such as *crime investigation*), but not its direct cause, should be labelled ABOUT by Rule 2.1.6.3.

As only one participant is mentioned, there is no need to establish its degree of agentivity.

Rule 2.1.5.2 *The compound is associated with a characteristic event in which N1/N2 and N2/N1 are participants, N1/N2 is more agentive than N2/N1, and N1/N2 is not an ACTOR.*

For example:

English: *rice cooker*(2), *tear gas*(2), *blaze victim*(1).

Dutch: *cadeaubon*(2), *worstmachine*(2).

Afrikaans: *traangas*(2), *voedselverwerker*(2).

The directionality of the relation is determined by the more agentive participant in the hierarchy given in Rule 1.5: *cheese*_O *knife*_I (INST2), *wine*_O *vinegar*_R (INST1), *wind*_A *damage*_R (INST1), *human*_O *virus*_A (INST1). Sometimes it may be difficult to distinguish Agents from Instruments (*gun wound*) or Objects from Results (*blaze victim*) – this is not important so long as it is possible to identify which participant is more agentive.

In some cases, it may not be clear what the exact underlying event is, but the more agentive participant may still be identified – a *transport system* is a system that in some way provides or manages transport, but it is nonetheless clear that the appropriate label is INST2. In other cases, where both participants affect each other, it may be less clear which is more agentive – *motor oil* can be construed as oil that lubricates/enables the function of the engine or as oil the engine uses. Likewise *petrol motor*, *computer software*, *electron microscope*. At least where the relation is between a system or machine and some entity it uses to perform its function, the former should be chosen as more agentive. Hence *motor oil* is INST1, *petrol motor* is INST2, and so on.

As in Rule 2.1.5.1, where one of the constituents is the location of the associated event, then IN is the appropriate label by Rule 2.1.3.1 or 2.1.3.3. If the more agentive participant meets the criteria for ACTOR status (2.1.4), then that label should be applied instead. If the interaction between the constituents is due to one being a part of the other (as in car engine), HAVE is the appropriate label by Rule 2.1.2.4. A border with ABOUT must be drawn in the case of psychological states and human activities whose cause or focus is N1. As described further under Rules 2.1.6.3, the criterion adopted is based on whether there is a direct causal link between N1 and N2 in the underlying event – a bomb can by itself cause *bomb terror* (INST1), but a *spider phobia* is not a reaction to any particular spider and is classed as ABOUT.

2.1.6 ABOUT

Rule 2.1.6.1 *N1/N2's descriptive, significative or propositional content relates to N2/N1.*

For example:

English: *fairy tale*(2), *flower picture*(2), *tax law*(2), *exclamation mark*(2), *film character*(2), *life principles*(2), *sitcom family*(1).

Dutch: *vakjargon*(2), *contactstoornis*(2), *praktijktheorie*(2), *vakdeskundigheid*(2).

Afrikaans: *mensekennis(2)*, *balletmusiek(2)*.

In English, a lot of speech acts belong to this category. Direction 2 is a lot more prominent with this rule. Properties and attributes that seem to have a descriptive or subjective nature are still to be labelled HAVE by Rule 2.1.2.3 – *street name* and *music loudness* are HAVE1.

Rule 2.1.6.2 *N1/N2 is a collection of items whose descriptive, significative or propositional content relates to N2/N1 or an event that describes or conveys information about N2/N1.*

For example:

English: *history exhibition(2)*, *war archive(2)*, *science lesson(2)*.

Dutch: *tijdreeks(2)*, *muziekbibliotheek(2)*.

Afrikaans: *kunsuitstalling(2)*, *musiekversameling(2)*.

Rule 2.1.6.3 *N1/N2 is a mental process or mental activity focused on N2/N1, or an activity resulting from such.*

For example:

English: *crime investigation(2)*, *science research(2)*, *research topic(1)*, *exercise obsession(2)*, *election campaign(2)*, *football violence(2)*, *holiday plan(2)*.

Dutch: *darmonderzoek(2)*, *plantenobsessie(2)*.

Afrikaans: *taalnavorsing(2)*, *selfondersoek(2)*.

In the case of activities, N1/N2 cannot belong to any of the participant categories given under Rule 1.5; rather it is the topic of or motivation for N2/N1. The sense of causation in, for example, *oil dispute* is not direct enough to admit an INST classification – the state of the oil supply will not lead to an oil dispute without the involved parties taking salient enabling action. In the case of emotions, there is also a risk of overlapping with INST; *bomb terror* is INST and *bomb dislike* is classed as ABOUT, but examples such as *bomb fear* are less clearcut. A line can be drawn whereby immediate emotional reactions to a stimulus are annotated INST, but more permanent dispositions are ABOUT. In the case of bomb fear, the relation must be identified from context. Problems (*debt problem*) and crises (*oil crisis*) also belong to this category, as they are created by mental processes.

Rule 2.1.6.4 *N1/N2 is an amount of money or some other commodity given in exchange for N2/N1 or to satisfy a debt arising from N2/N1.*

For example:

English: *share price*(2), *printing charge*(2), *income tax*(2).

Dutch: *olieprij*s(2), *loonarbeid*(1), *gokbedrag*(2).

Afrikaans: *goudprys*(2), *boedelbelasting*(2), *petrolprys*(2), *bankkoste*(2).

N2/N1 is not the giver or recipient of N1/N2 – an *agency fee* would be INST under the interpretation fee_I paid to an agency_O – but the thing exchanged or the reason for the transaction.

2.1.7 REL

Rule 2.1.7.1 *The relation between N1 and N2 is not described by any of the above relations but seems to be produced by a productive pattern.*

For example:

English: *Baker Street*, *sodium chloride*,

Dutch: *Vaarttheater*, *Plataanlei*, *waterstofcarbonaat*, *adjudant-onderofficier*.

Afrikaans: *Akkerlaan*, *waterstofkarbonaat*.

A compound can be associated with a productive pattern if it displays substitutability. If both of the constituents can be replaced by an open or large set of other words to produce a compound encoding the same semantic relation, then a REL annotation is admissible. For example, the compound *reading skill* (in the sense of degree of skill at reading) is not covered by any of the foregoing categories, but the semantic relation of the compound (something like ABILITY) is the same as that in *football skill*, *reading ability* and *learning capacity*. This contrasts with an idiosyncratic lexicalised compound such as *home secretary* (= LEX), where the only opportunities for substitution come from a restricted class and most substitutions with similar words will not yield the same semantic relation. Another class of compounds that should be labelled REL are names of chemical compounds such as *carbon dioxide* and *sodium carbonate*, as they are formed according to productive patterns. There are also several special cases that receive the REL tag. Take a look at Rule 1.4 for the descriptions.

2.1.8 LEX

Rule 2.1.8.1 *The meaning of the compound is not described by any of the above relations and it does not seem to be produced by a productive pattern.*

For example:

English: *turf accountant*, *monkey business*.

Dutch: *loftrompet*, *prins-gemaal*, *spierbundel*.

Afrikaans: *spierpaleis*.

These are noncompositional in the sense that their meanings must be learned on a case-by-case basis and cannot be identified through knowledge of other compounds. This is because they do not have the property of substitutability - the hypothetical compounds *horse business* or *monkey activity* are unlikely to have a similar meaning to *monkey business*. LEX also applies where a single constituent has been idiosyncratically lexicalised as a modifier or head such as *X secretary* meaning ‘minister responsible for X’.

2.1.9 UNKNOWN

Rule 2.1.9.1 *The meaning of the compound is too unclear to classify.*

Some compounds are simply uninterpretable, even in context. This label should be avoided as much as possible but is sometimes unavoidable.

2.2 Noncompounds

2.2.1 MISTAG

Rule 2.2.1.1 *One or both of N1 and N2 have been mistagged and should not be counted as (a) common noun(s).*

For example:

English: *fruity bouquet* (N1 is an adjective), *London town* (N1 is a proper noun).

Dutch: *Juratijdperk* (N1 is a proper noun), *voortuin* (N1 is a preposition), *hoofdbewaker* (N1 is adjective-like).

Afrikaans: *geelwortel* (N1 is an adjective), *hoofkok* (N1 is adjective-like), *afhaal* (N is a verb), *Paryspad* (N1 is a proper noun).

In the case of *blazing fire*, N1 is a verb, so this is also a case of mistagging; in superficially similar cases such as *dancing teacher* or *swimming pool*, however, the -ing form can and should be treated as a noun. The annotator must decide which analysis is correct in each case – a *dancing teacher* might be a teacher who is dancing (MISTAG) in one context, but a teacher who teaches dancing (ACTOR) in another context. Certain modifiers might be argued to be nouns but for the purposes of annotation are stipulated to be adjectives. Where one of *assistant*, *key*, *favourite*, *deputy*, *head*, *chief* or *fellow* appears as the modifier of a compound in the data, it is to be considered mistagged. This only applies when

these modifiers are used in adjective-like senses – *key chain* or *head louse* are clearly valid compounds and should be annotated as such.

2.2.2 NONCOMPOUND

Rule 2.2.2.1 *The extracted sequence, while correctly tagged, is not a 2-noun compound.*

There are various reasons why two adjacent nouns may not constitute a compound:

1. An adjacent word should have been tagged as a noun, but was not.
2. The modifier is itself modified by an adjacent word, corresponding to a bracketing $[[X\ N1]\ N2]$. For example: $[[real\ tennis]\ club]$, $[[Liberal\ Democrat]\ candidate]$, $[[five\ dollar]\ bill]$. However compounds with conjoined modifiers such as *land and sea warfare* and *fruit and vegetable seller* can be treated as valid compounds so long as the conjunction is elliptical (*land and sea warfare* has the same meaning as *land warfare* and *sea warfare*). Not all conjoined modifiers satisfy this condition – a *salt and pepper beard* does not mean a beard which is a *salt beard* and a *pepper beard*, and the sequence *pepper beard* is a NONCOMPOUND.
3. The two words are adjacent for other reasons. For example: ‘the *question politicians* need to answer’, structureless lists of words.
4. The modifier is not found as a noun on its own, because it would not appear in the dictionary. For example: *multiparty election*, *smalltown atmosphere*.

ACKNOWLEDGEMENT

This protocol was developed for research on automatic compound processing. Automatic Compound Processing (AuCoPro) is a mutual project by research groups of the North-West University (Potchefstroom, South Africa), the University of Antwerp (Belgium) and Tilburg University (The Netherlands). The University of Antwerp deals mainly with the compound semantics subproject, Tilburg University deals mainly with compound splitting. North-West University works on the Afrikaans aspects of both subprojects.

This research was co-funded by a joint research grant of the Nederlandse Taalunie (Dutch Language Union) and the Department of Arts and Culture (DAC) of South Africa and a grant of the National Research Foundation (NRF) (grant number 81794).