Guidelines for the Annotation of Coordinators and Subordination Boundaries (FIRST-287607:WP7:3:Version 1)

Richard Evans, Emma Franklin, and Zoe Harrison
Research Group in Computational Linguistics,
University of Wolverhampton,
United Kingdom,
R.J.Evans@wlv.ac.uk

July 12, 2013

Abstract

The focus of these guidelines is the manual annotation of coordinators and subordination boundaries, which comprise conjunctions, complementisers, wh-words, punctuation, and adjacent pairs of these words and punctuation symbols. In addition to guidelines, the document presents information on the annotation scheme used to develop training data for the automatic classification of coordinators and subordination boundaries.

1 Introduction

Syntactic simplification can be regarded as the process of converting syntactically complex sentences into one or more simpler sentences. For the purpose of syntactic simplification, complexity is considered to involve coordination and subordination. These phenomena are signalled by the explicit occurrence of coordinators and subordination boundaries in text.

This document is structured as follows. Section 2 presents the annotation scheme, the markables, and the classes to which signs of syntactic complexity should be assigned. Sections 3-4 introduces the notion of coordination and provides examples of different classes of these coordinators and subordination boundaries. Sections 5–6 present examples of additional uses of ambiguous signs of syntactic complexity. Section 7 provides examples of the annotation of three English sentences while Section 8 discusses cases of uncertainty that are likely to cause inconsistency in the annotation process. Section 9 is an appendix intended to facilitate interpretation of the class labels by annotators.

2 The annotation scheme

This section presents general information about the things to be classified (Section 2.1) and the classes to which they may belong (Sections 3 and 4).

2.1 Markables: Coordinators and Subordination Boundaries

In this project, the markables are items in the text that serve either as coordinators or as subordination boundaries. Coordination and subordination are always considered to link two conjoins. It is common for three or more conjoins to be linked within a sentence, but when considered individually, each case of coordination or subordination links just two elements. The task of the annotator is to classify each coordinator according to the type of constituents that it links. In the FIRST project, coordinators comprise the sets:

- conjunctions: {*and*, *but*, *or*};
- punctuation: {,, ;, :} (the comma and semicolon may have either subordinating or coordinating functions);
- punctuation-conjunction pairs: {*, and*, *; and*, *: and*, *, but*, *; but*, *: but*, *, or*, *; or*, *: or*,

In the project, subordination boundaries comprise the sets:

- complementiser: {*that*};
- wh-words: {*what*, *when*, *where*, *which*, *while*, *who*};
- punctuation: {,, ;, :} (the comma and semicolon may have either subordinating or coordinating functions);
- punctuation-conjunction pairs: {*, and*, *; and*, *: and*, *, but*, *; but*, *: but*, *, or*, *; or*, *: or*,
- punctuation-complementiser pairs: {*, that*, *; that*, *: that*},
- punctuation-wh-word pairs: {*, what*, *; what*, *: what*, *, when*, *; when*, *: when*, *, where*, *; where*, *: where*, *, which*, *; which*, *: which*, *, while*, *; while*, *: while*, *, who*, *; who*, *: who*}, most frequently indicating subordination, but ambiguous due to the various functions of the punctuation symbol.

Annotators are expected to assign markables to classes that provide information on the syntactic category and projection level of the constituents that they coordinate (for coordinators) or the subordinated constituents that they bound (for subordination boundaries). In the current guidelines, these constituents are also referred to as *conjoins* [Quirk et al., 1985].

3 Coordination

Coordination is a paratactic relationship that holds between constituents at the same level of syntactic structure (i.e. between "sister" nodes in the syntactic tree). Coordination usually occurs at the lexical or phrasal level, but can also occur between clauses, prefixes, and intermediate categories such as $\bar{\text{N}}$.

There are considered to be 21 types of coordination in the annotation scheme, as listed in Sections 3.1-3.21, with examples of their use. Throughout this document, when examples are provided, potential coordinators are marked in square brackets and the linked constituents (conjoins) are underlined. Where appropriate, the location of elided elements is indicated using ϕ. Occurrences of ϕ may be co-indexed.

3.1 CLN: Head Nouns

(1) It is vital that truth [and] justice are seen to be done.

(2) There was no history of any animosity between him [and] Gillian or Mr Brown or Mr Smith.

(3) Nursing care for Ms Coughlan, who requires round-the-clock help, and a fellow resident who is doubly incontinent; immobile and unable to communicate, was the NHS's responsibility Help with feeding [and] bathing would presumably be part of the social services' care package.

3.2 CIN: Intermediate Nominal Projections (N-Bar)

This type of coordination can be distinguished from CLN because an adjective in one conjoin should not modify the head noun of the other conjoin. It can be distinguished from CMN1 because the second conjoin cannot have an initial specifier (e.g. determiner), instead the two conjoins share the specifier of the first.

(4) My property, house, vehicle, savings ... as well as my private company [and] clients.

(5) A WOMAN paralysed after a road accident has lost her test case in the Court of Appeal to force the state to pay for all the long-term nursing care of the chronically sick [and] disabled. (may be CLN if *chronically* modifies *disabled*)

(6) Has this case not revealed a dark side to Bruce Grobbelaar, a degree of intrigue, fondness for deception [and] interest in money? (? N-bars or just NPs? Does *interest in money* need *an* as a specifier rather than *a*?)

3.3 CMN1: Noun Phrases (NP)

(7) He also FLAUNTED a £30,000 watch, DRIPPED with gold and other jewellery and BOASTED he ran four luxury motors - a £50,000 Mercedes 320 SL[,] a Porsche, a BMW and a Saab turbo.

(8) "Occasionally I ate pasty and chips [or] jumbo sausage with chips, but mainly it was just chips," he said.

(9) Mr Justice Forbes told the pharmacists that both Mr Young [and] his girlfriend, Collette Jackson, 24, of Runcorn, Cheshire, had been devastated by the premature loss of their son.

(10) Light nursing care, such as giving tablets [and] helping with feeding, would probably fall on social services.[1]

3.4 CMN2: NPs coordinated to specify a superordinate NP

In future, this class may be considered obsolete and instances reclassified as occurrences of either CLN or COMBINATORY. There is currently little justification for CMN2.

(11) Three months later the Nestlé-owned plant was deluged with complaints from customers who had found glass in minced beef [and] chicken pancakes.

(12) They believe his lieutenants bought the drugs in London and Birmingham and that he used boys as young as 12 to ferry drugs on mountain bikes around Nottingham's St Ann's [and] Meadows estates.

(13) Mr Clarke was at that time unaware Noye had fatally stabbed Stephen, 21, at the M25 [and] M20 interchange near Swanley, Kent.

3.5 CMN3: Obsolete class for use in a restricted genre (assessment) and a restricted domain (clinical)

3.6 CMN4: NPs in which the head noun of one or more conjoins is elided

(14) 'This case is not about whether GM crops are a good ϕ_i [or] a bad thing$_i$,' he said.

(15) Brent Magistrates' Court was told he was subsequently convicted of six counts$_i$ of theft [and] one ϕ_i of attempted theft last June.

(16) Her spinster daughter had a stab wound between the shoulder blades, a stab wound$_i$ on each side of her neck [and] six ϕ_i in her chest.

3.7 CLAdv: Head Adverbs

(17) "Anaïs is doing very[,] very well."

(18) "His confidence was completely [and] utterly gone."

(19) 'I have struggled for nearly three years to reach this point and it has taken every thing inside me mentally, physically[,] emotionally to get here.

[1] This is an example of coordinated nominal -ing participle clauses, which could be mistaken for verb phrases and classified as CMV1.

3.8 CMAdv: Adverbial Phrases

(20) 'I suspect I will bid a fond goodbye before I get too stale [and] before I get too frustrated,' said Edmonds.

(21) He claims to have paid Mr Hamilton up to £110,000, either directly [or] through the lobbyist Ian Greer, in return for parliamentary services.

(22) "One's heart goes out to the parents of the boy who died so tragically [and] so young."

3.9 CMA1: Adjectival Phrases

(23) "Yolanda was painfully shy [and] quite unable to join in the friendly banter in the classroom," she said.

(24) By day he was polite, softly spoken [and] conscientious, working mostly in minicabs or as a truck driver.

(25) Mr Carman said that the statement was not only false [but] also accepted by both newspapers to be untrue.

3.10 CMA2: Obsolete class for use in a restricted genre (assessment) and a restricted domain (clinical)

3.11 CLA: Head Adjectives

(26) "He had a stable [and] loving family."

(27) "Only the BBC can give new comedy the patient [and] pressure-free environment it needs to flourish," said Mr Salmon.

(28) "Everyday, health [and] local authorities are squabbling over who pays for long-term care, and older people are caught in the middle.

3.12 CPA: Adjectival Prefixes

(29) A heavy police presence kept watch on rival pro [and] anti-Pinochet groups chanting outside.

(30) Ultrasonography of the abdomen shows cholelithiasis with intra- [and] extra-hepatic biliary dilation.

(31) Technicolor's Creative Services in London provide pre [and] post production services in film, broadcast, 3D, sound design and subtitling.

3.13 CLP: Head Prepositions

(32) A court martial found Company Sgt-Major Michael Gleave, 39, not guilty of six charges of making racial remarks to [or] about Pte Roy Carr, 26, and two of ill treatment.

(33) A High Court judge yesterday condemned the Government for misleading parents about its policy on assisted places in promises before [and] after the general election.

(34) Banfield carried out a sustained and systematic year-long campaign of abuse, mostly while on duty in uniform in [or] near Parkside police station, Cambridge, where he served as a custody sergeant.

3.14 CMP: Prepositional Phrases

(35) "But the melancholy experience of the courtrooms [and] of life is that people have a good character in some respects and not in others."

(36) BA described the decision as "wrong in fact [and] in law" and said it would appeal.

(37) Denning was charged in Prague with having sexually assaulted an array of boys, some as young as 12[, and] of being the head of a paedophile ring which included two Frenchmen and an American.

3.15 CMP2: Prepositional Phrases with elision of the head preposition of one or more conjoins

This class may also be withdrawn. Although two instances are presented below, their classification is debateable. The remaining instances annotated in the corpus can be considered occurrences of CLN or CMN1.

(38) Asked why, he was heard to say: 'That's for$_i$ me to know [and] ϕ_i you to find out.'

(39) In the autumn of 1995, five current and former staff filed written allegations, accusing Francis of dealing in heroin and cocaine, of$_i$ trading in firearms [and] ϕ_i supplying women for prostitution.

3.16 CLV: Head Verbs

(40) 'They were kicking [and] punching each other.'

(41) Drummond was jailed for three months concurrently on each of six charges of wilfully killing, taking [and] mistreating badgers.

(42) He overpowered [and] bound Janice Sheridan with adhesive tape after stabbing her in the back and then butchered her mother.

3.17 CMV1: Verb Phrases

(43) "Everyone has been talking about it [and] praying for her."

(44) "I will be going to see it in a couple of days [and] am really looking forward to celebrating with them then."

(45) McKay had been on the payroll five years[, but] feared he would go on a 'last-in, first out' basis, said Mr Sloan.

3.18 CMV2: Verb Phrases in which the head of one conjoin is elided

(46) But Justice Kay added: "It is a sorry state of affairs when Mr Blunkett has to explain away$_i$ his own letters as mistaken and unclear [and] ϕ_i a statement by the Prime Minister as an incorrect representation of policy, taken out of context".

(47) 'The price discussed with John Holmes was$_i$ £20,000 for Paul Paterson [and] ϕ_i £20,000 for his wife.

(48) They sent$_i$ one surveillance team to follow the London suppliers as they drove up the motorway[, and] ϕ_i another team to Francis's flat in the city centre.

3.19 CCV: Clauses

(49) "What I have really been looking forward to is a Cornetto [and] at last I have had one."

(50) Brian and I wanted to live here[,] we wanted to emigrate, and if I could, I would.

(51) A paraplegic car crash victim who needed specialist care [and] whose condition could change rapidly would certainly get free care.

3.20 CLQ: Quantifiers

(52) And after the jury had been out for six [and] a half days, he was convicted only of grievous bodily harm.

(53) Anyone with assets of £10,000 [or] more, including the value of their homes, has to contribute towards the cost of social services run homes, while those with £16,000 get no help at all.

(54) I was dragged down to the floor and hit on the back of the legs three [or] four times.

3.21 COMBINATORY: Unsplittable

(55) McKay was told he was third on the list to go, following the "last-in[,] first-out" policy.

(56) Mrs Druhan's case was first raised seven years ago by the Channel 4 programme Trial [and] Error, for which the result is a triumph.

(57) The 62-year-old ex-SAS man, who was awarded the Military Medal by the Queen, was left high [and] dry in the witness box after the judge, Mr Justice Morland, walked out, fed up with his unsolicited outbursts.

4 Subordination

Subordination is a hypotactic relationship between constituents at different levels of syntactic structure (between subordinate/superordinate phrases). Subordination holds between phrases or clauses, rather than between lexical or intermediate constituents. There is no expectation that the syntactic categories of these phrases will match. Subordinate constituents provide additional information about the head of the superordinate constituent.

There are currently considered to be nine types of subordinated constituent. Each type may be bounded to the immediate left or right by a markable.[2] In many cases, the rightmost boundary of a subordinated constituent is marked by means of a comma or other punctuation mark. However, in many cases annotation is complicated by the facts that:

1. multiple rightmost boundaries can be coincident;
2. coordinators can serve as the rightmost boundary of subordinated constituents;
3. sentence boundaries can serve as the rightmost boundary of subordinated constituents; and
4. occasionally, the rightmost boundary of a subordinated constituent is not explicitly marked.

4.1 Leftmost Boundaries

4.1.1 SSMAdvP: Adverbial Phrase

(58) 'Specialist nursing care' remains the responsibility of the NHS[,] however, and is free.

(59) He goes around putting two fingers up to everyone else[,] usually quite literally.

(60) "I have struggled for nearly three years to reach this point and it has taken everything inside me[,] mentally, physically and emotionally, to get here.

4.1.2 SSCCV: Clause

The class denoting leftmost boundaries of subordinate clauses is by far the most common annotated in the METER corpus. SSCCV currently includes several different types of clause, including nominal *that*-clauses; adverbial *when*-clauses; non-finite clauses (including *ed*-participle clauses, infinitive clauses (bare infinitive and *to*-infinitive clauses));[3] some verbless clauses;[4]; *ing*-clauses; and *ed*-clauses. No au-

[2]The exceptions to this are tag questions which are usually bounded by question marks to the right, rather than the markables included in this annotation scheme.

[3]Considered SSCCV when the subordinator is a *wh*-word, SSMV when the subordinator is a punctuation mark.

[4]In many cases, the leftmost boundary of a verbless clause is better annotated as an instance of SMAdv or SMP.

tomatic classifier of subordination boundaries has yet been implemented, but the results of this may motivate division of class SSCCV into various subclasses.

(61) "And Oliver is a man [who] crumbled and buckled."

(62) "That's simply not true," said Grobbelaar[,] who is suing The Sun for libel over allegations of match-fixing.

(63) The court heard [that] Hobbs struck him with the blunt end of an axe and tied him on the bed next to Mr Brown.

4.1.3 SSMV: Verb Phrase

(64) "He kept Mr Brown prisoner until the next day[,] sleeping in the house that night."

(65) And he claimed he was engaged in dirty tricks[,] keeping a book of people who used to collect envelopes of cash.

(66) Michael Lawson QC[,] prosecuting, said Littlebury bought a 12-bore shotgun after going on the shooting course.

4.1.4 SSMP: Prepositional Phrase

(67) A month later[,] on January 3 1983, Barwell was back on their patch.

(68) And Kattab[,] of Eccles, was not aware of the different dilutions of chloroform used to make peppermint oil.

(69) He stressed he accepted that Hitler[,] "as head of state and of the government", was responsible for the Holocaust.

4.1.5 SSMN: Noun Phrase

(70) A son[,] Tyrell, followed in 1970.

(71) Actor Alec Baldwin will play Mr Conductor[,] a new character specially created for Thomas And The Magic Railroad.

(72) Adrian Shaw, a lorry driver from Malvern, Worcestershire, and his employer[,] Edward Gilder, of Cheltenham, Gloucestershire, deny two charges of animal cruelty.

4.1.6 SSCM: Reported Speech

(73) 'He said[,] "You kill him."

(74) "Always remember," he said[,] "that The Beatles were a rock'n'roll band and that's why we were so good for so long, if that's not too immodest."

(75) The father claimed[:] 'The police have victimised him since he was nine years old, and the schools have let him down.'

4.1.7 SSMA: Adjective Phrase

(76) "It is wonderful news," said his mother Patricia[,] 61.

(77) 'There is no expectation that Irish people[,] alive or dead, can achieve justice in British courts,' he said.

(78) A jury at Bristol Crown Court heard that Mr Guscott[,] furious at having to brake, decided to 'teach Mr Jones a lesson'.

4.1.8 SSMI: Interjection

So far, of 5 220 leftmost subordination boundaries, only two instances of SSMI have been annotated:

(79) 'No[,] my lord,' conceded Collins's barrister Robert Howe.

(80) 'I have told you before I am a father who has lost his son and I have the right to do anything to find out how I lost my son and[,] please, I have asked you several times not to capitalise on my grief.'

4.1.9 STQ: Tag Question

(81) 'That's not a very attractive position[,] is it?'

(82) "You won't go near women again[,] will you?"

(83) 'They've done a pretty good job[,] haven't they?' he'd told me earlier, looking around.

4.2 Rightmost Boundaries

4.2.1 ESMAdvP: Adverbial Phrase

(84) "However[,] his account left me troubled."

(85) Earlier[,] Judge Caroline Simpson had expressed her personal sympathy for Mr Hagland's friends and relatives.

(86) But, yesterday[,] the eagerly anticipated entertainment came to an abrupt end when Marco Pierre White, the enfant terrible of British cooking, claimed a legal victory and the right to continue trading at Titanic, his £2m restaurant in London's West End.

4.2.2 ESCCV: Clause

(87) As she emerged from the cells[,] Ms Druhan exclaimed: "I'm free."

(88) Detectives were shocked at the severity of the attacks on the women[,] several of whom said the assailant was Scottish because he referred to them as "lassie".

(89) His order forced Miss Parker, who was then 16[,] to be detained at a clinic and undergo life-saving treatment for her eating disorder.

4.2.3 ESMV: Verb Phrase

(90) Baggs, defending[,] said the offence was "out of character".

(91) Comparing her case with the heart transplant order[,] Chaye said Child M would one day be thankful for the decision.

(92) "Being put into a psychiatric ward with people with long-term mental illnesses who are shaking with the drugs they are taking[,] there's no way you can feel normal and be OK with yourself," she told BBC TV's That's Esther programme with Esther Rantzen.

4.2.4 ESMP: Prepositional Phrase

(93) In Dundee, for example[,] there are hardly any council residential homes.

(94) After more than a week's deliberation[,] an Australian jury did not even find the drunken thug guilty of manslaughter.

(95) Former anorexic Chaye Parker, from Stoke Newington, London[,] was ordered to be detained if she resisted treatment.

4.2.5 ESMN: Noun Phrase

(96) A second man, Sean Cushman, 25[,] was convicted of being an accessory and also awaits sentence.

(97) George Carman, QC, defending the newspaper[,] said the tapes confirmed what The Sun had been told earlier about Grob by his former business partner Chris Vincent.

(98) His friends, Mark Picard, and Earl Petrie, both 24 and both of Kingston[,] were jailed for three months and 12 months respectively for their parts in the attack on Mr Lee and a man trying to help him.

4.2.6 ESCM: Reported Speech

(99) "I thought he was going to hit me[,]" said Bennett.

(100) "It was real shock for all of us[,]" one said yesterday.

(101) "He stayed there for about two-and-a-half hours[,]" said Miss Barry "He drank three pints of Kronenberg."

4.2.7 ESMA: Adjective Phrase

(102) A court martial took four hours to decide Sergeant Major Gleave, 39[,] was not guilty of eight charges.

(103) "Inquisitive by nature[,] he lacked the experience to deal with a situation into which his curiosity had led him," they said.

(104) Andrew Hawkins, 42, of Ham Farm Lane, Bristol[,] admitted 14 specimen charges under trading standards laws, but Exeter Crown court was told that he had altered the odometers on hundreds of cars in Britain's worst car clocking case.

4.2.8 ESMI: Interjection

(105) 'I thought 'Yes[,] I will remember'.'

(106) 'Ladies and gentlemen[,] Paul McCartney and his band,' announced an unseen presenter to the 300-strong audience - and there he was, back where he began, doing what he always did best, playing rock 'n' roll as he raced straight into an old Big Joe Turner song, Honey Hush.

(107) 'I have told you before I am a father who has lost his son and I have the right to do anything to find out how I lost my son and, please[,] I have asked you several times not to capitalise on my grief.'

5 Examples of Special Uses of Markables (Which Currently Cannot be Adequately Classified / Simplified Using Our Current Algorithm)

(108) "Name something that is currently on BBC1 that gets people excited [and] talking about it.

(109) "I'm quite happy to abandon [that] specific point" he said.[5]

(110) I intended to say: 'You went through a red light', only I don't remember whether I said it [or] not as the next thing I knew I was being grabbed by the defendant.[6].

It should be noted that multiple nesting of subordinate constituents is common (e.g. (111)). This means that various markables serve multiple functions. Annotators should assume that the boundaries of a conjoin also bound any subordinate constituent of that conjoin. As a result, ambiguous instances are annotated as the left or right boundaries of the most superordinate (larger) constituents as opposed to more subordinate ones. To illustrate, the indicated comma in (111) would be annotated as ESMA, the rightmost boundary of the subordinated adjectival phrase whose leftmost boundary is the first comma appearing in the sentence. The indicated comma also signals the rightmost boundary of a subordinated prepositional phrase, but this is itself subordinate to the adjectival phrase.

(111) David Machin, 40, of Wakefield Yorkshire[,$_{ESMA}$] is charged with outraging public decency.

[5] *That* used as a determiner rather than as a complementiser/subordinator.

[6] Ill-assorted conjoins.

Appendix **??** presents examples of sentences containing subordinated constituent boundaries.

6 Currently Unclassifiable / Unsimplifiable by means of our current algorithm

The label SPECIAL is used to denote instances of this class, which includes cases of ill-assorted coordination, combinatory coordination, the use of *that* as a determiner or anaphor rather than a complementiser, and more. Sentences containing such instances are not expected to be simplified automatically, and will be ignored by the software developed in the project.

Examples of sentences containing this type of potential coordinator are provided in Appendix **??**. Some of these cases are also discussed in Section 8.

7 Example Annotation

In this Section, the annotation of every coordinator and subordination boundary in three example sentences is presented.

7.1 Example 1

Considering each markable in turn, Sentence (112) is annotated in the following way. The first comma follows a subordinated prepositional phrase, and serves as its rightmost boundary:

(112) As researcher for the programme[$_{ESMP}$,] Ms Price had arranged for the guests to appear on the show and the article went on to allege that not only were they fakes but that Ms Price had known this and had deliberately deceived her employers and viewers.

The first conjunction links two clauses in coordination:

> As researcher for the programme, Ms Price had arranged for the guests to appear on the show [$_{CCV}$ and] the article went on to allege that not only were they fakes but that Ms Price had known this and had deliberately deceived her employers and viewers.

The second conjunction links two subordinate clauses in coordination:

> As researcher for the programme, Ms Price had arranged for the guests to appear on the show and the article went on to allege that not only were they fakes [$_{CCV}$ but] that Ms Price had known this and had deliberately deceived her employers and viewers.

The third conjunction links two predications (verb phrases) in coordination:

As researcher for the programme, Ms Price had arranged for the guests to appear on the show and the article went on to allege that not only were they fakes but that Ms Price had known this [$_{CMV1}$ and] had deliberately deceived her employers and viewers.

The fourth conjunction links two nouns in coordination:

As researcher for the programme, Ms Price had arranged for the guests to appear on the show and the article went on to allege that not only were they fakes but that Ms Price had known this and had deliberately deceived her employers [$_{CLN}$ and] viewers.

The fully annotated sentence is:

As researcher for the programme[$_{ESMP}$,] Ms Price had arranged for the guests to appear on the show [$_{CCV}$ and] the article went on to allege that not only were they fakes [$_{CCV}$ but] that Ms Price had known this [$_{CMV1}$ and] had deliberately deceived her employers [$_{CLN}$ and] viewers.

7.2 Example 2

The first comma in Sentence (113) introduces a subordinate prepositional phrase that provides additional information about *The Grants*. It is annotated as:

(113) The Grants[$_{SSMP}$,] of The Old Coach House, Sully, in the Vale of Glamorgan, South Wales, denied three charges of cruelty from 1994 to 1997: cruelty to a child by regular violent attacks; cruelty by failing to provide adequate food, clothing and accommodation; and cruelty by constantly referring to him in a derogatory fashion.

The second comma introduces a subordinate noun phrase, providing information to specify the location of *The Old Coach House* (and *The Grants*) more precisely:

The Grants, of The Old Coach House[$_{SSMN}$,] Sully, in the Vale of Glamorgan, South Wales, denied three charges of cruelty from 1994 to 1997: cruelty to a child by regular violent attacks; cruelty by failing to provide adequate food, clothing and accommodation; and cruelty by constantly referring to him in a derogatory fashion.

The third comma introduces a subordinate prepositional phrase, further specifying the location of *Sully* and *The Old Coach House*:

The Grants, of The Old Coach House, Sully[$_{SSMP}$,] in the Vale of Glamorgan, South Wales, denied three charges of cruelty from 1994 to 1997: cruelty to a child by regular violent attacks; cruelty by failing to provide adequate food, clothing and accommodation; and cruelty by constantly referring to him in a derogatory fashion.

The fourth comma introduces another subordinate noun phrase, continuing this specification:

> The Grants, of The Old Coach House, Sully, in the Vale of Glamorgan[$_{SSMN}$,] South Wales, denied three charges of cruelty from 1994 to 1997: cruelty to a child by regular violent attacks; cruelty by failing to provide adequate food, clothing and accommodation; and cruelty by constantly referring to him in a derogatory fashion.

The fifth comma has multiple functions, serving as the rightmost boundary of several subordinate phrases that provide information on the subject of the sentence: *South Wales*; *in the Vale of Glamorgan, South Wales*; *Sully, in the Vale of Glamorgan South Wales*; and *of The Old Coach House, Sully, in the Vale of Glamorgan, South Wales*. As noted in Section 4, the choice of annotation is based on the syntactic category and projection level of the most superordinate (largest) of these subordinate constituents:

> The Grants, of The Old Coach House, Sully, in the Vale of Glamorgan, South Wales[$_{ESMP}$,] denied three charges of cruelty from 1994 to 1997: cruelty to a child by regular violent attacks; cruelty by failing to provide adequate food, clothing and accommodation; and cruelty by constantly referring to him in a derogatory fashion.

The colon introduces a list of noun phrases. Such lists are considered structurally subordinate:

> The Grants, of The Old Coach House, Sully, in the Vale of Glamorgan, South Wales, denied three charges of cruelty from 1994 to 1997[$_{SSMN}$:] cruelty to a child by regular violent attacks; cruelty by failing to provide adequate food, clothing and accommodation; and cruelty by constantly referring to him in a derogatory fashion.

The first semicolon links two noun phrases in coordination:

> The Grants, of The Old Coach House, Sully, in the Vale of Glamorgan, South Wales, denied three charges of cruelty from 1994 to 1997: cruelty to a child by regular violent attacks[$_{CMN1}$;] cruelty by failing to provide adequate food, clothing and accommodation; and cruelty by constantly referring to him in a derogatory fashion.

The sixth comma links two nouns in coordination:

> The Grants, of The Old Coach House, Sully, in the Vale of Glamorgan, South Wales, denied three charges of cruelty from 1994 to 1997: cruelty to a child by regular violent attacks; cruelty by failing to provide adequate food[$_{CLN}$,] clothing and accommodation; and cruelty by constantly referring to him in a derogatory fashion.

The first conjunction links two nouns in coordination:

The Grants, of The Old Coach House, Sully, in the Vale of Glamorgan, South Wales, denied three charges of cruelty from 1994 to 1997: cruelty to a child by regular violent attacks; cruelty by failing to provide adequate food, clothing [$_{CLN}$ and] accommodation; and cruelty by constantly referring to him in a derogatory fashion.

The final markable links two noun phrases in coordination:

The Grants, of The Old Coach House, Sully, in the Vale of Glamorgan, South Wales, denied three charges of cruelty from 1994 to 1997: cruelty to a child by regular violent attacks; cruelty by failing to provide adequate food, clothing and accommodation[$_{CMN1}$; and] cruelty by constantly referring to him in a derogatory fashion.

The fully annotated sentence is:

The Grants[$_{SSMP}$,] of The Old Coach House[$_{SSMN}$,] Sully[$_{SSMP}$,] in the Vale of Glamorgan[$_{SSMN}$,] South Wales[$_{ESMP}$,] denied three charges of cruelty from 1994 to 1997[$_{SSCCV}$:] cruelty to a child by regular violent attacks[$_{CMN1}$;] cruelty by failing to provide adequate food[$_{CLN}$,] clothing [$_{CLN}$ and] accommodation[$_{CMN1}$; and] cruelty by constantly referring to him in a derogatory fashion.

7.3 Example 3

The first comma in Sentence (114) introduces a subordinate noun phrase which provides more information on *Adrian Shaw*. It is annotated as:

(114) Adrian Shaw[$_{SSMN}$,] a lorry driver from Malvern, Worcestershire, and his employer, Edward Gilder, of Cheltenham, Gloucestershire, deny two charges of animal cruelty.

The second comma is the leftmost boundary of a subordinated noun phrase providing more information about *Malvern*.

Adrian Shaw, a lorry driver from Malvern[$_{SSMN}$,] Worcestershire, and his employer, Edward Gilder, of Cheltenham, Gloucestershire, deny two charges of animal cruelty.

The third markable serves two functions. First, it is the rightmost boundary of the subordinate noun phrase bounded on the left by the second comma. Second, it links two noun phrases in coordination. In coordination, it is assumed that the boundaries of a conjoin also serve to bound any subordinate constituent of that conjoin. In light of this property, the annotation of the third markable reflects its second function: noun phrase coordination.

Adrian Shaw, a lorry driver from Malvern, Worcestershire[$_{CMN1}$, and] his employer, Edward Gilder, of Cheltenham, Gloucestershire, deny two charges of animal cruelty.

The third comma is the leftmost boundary of a subordinated noun phrase providing information about Adrian Shaw's employer:

> Adrian Shaw, a lorry driver from Malvern, Worcestershire, and his employer[$_{SSMN}$,] Edward Gilder, of Cheltenham, Gloucestershire, deny two charges of animal cruelty.

The fourth comma is the leftmost boundary of a subordinated prepositional phrase that provides additional information about the location of this employer:

> Adrian Shaw, a lorry driver from Malvern, Worcestershire, and his employer, Edward Gilder[$_{SSMP}$,] of Cheltenham, Gloucestershire, deny two charges of animal cruelty.

This information is further specified by the subordinate noun phrase introduced by the fifth comma.

> Adrian Shaw, a lorry driver from Malvern, Worcestershire, and his employer, Edward Gilder, of Cheltenham[$_{SSMN}$,] Gloucestershire, deny two charges of animal cruelty.

The final markable serves at the rightmost boundary of the subordinate noun phrase whose leftmost boundary is the fourth markable in the sentence (a comma):

> Adrian Shaw, a lorry driver from Malvern, Worcestershire, and his employer, Edward Gilder, of Cheltenham, Gloucestershire[$_{ESMN}$,] deny two charges of animal cruelty.

The fully annotated sentence is:

> Adrian Shaw[$_{SSMN}$,] a lorry driver from Malvern[$_{SSMN}$,] Worcestershire[$_{CMN1}$, and] his employer[$_{SSMN}$,] Edward Gilder[$_{SSMP}$,] of Cheltenham[$_{SSMN}$,] Gloucestershire[$_{ESMN}$,] deny two charges of animal cruelty.

8 Causes of Uncertainty and Inconsistency

Analysis of the annotated resources and inter-annotator agreement revealed seven main causes of uncertainty and inconsistency. The ability of annotators to recognise and adopt a consistent method to handle each case will facilitate the development of reliable annotated resources.

8.1 Coincident Subordination Boundaries

In many cases, a single markable serves simultaneously as the rightmost boundary of several subordinated constituents. Consider (115), which is annotated as (116).

(115) Graham Wilson, 25, of Grangetown, Middlesbrough, had admitted murdering Amanda Jane Fulcher, 21, a waitress, after persuading her to go to the house of a friend, Steven Staton, 20, who was jailed for six years for helping dispose of the body.

(116) Graham Wilson[,$_{SSMA}$] 25[,$_{SSMP}$] of Grangetown[,$_{SSMN}$] Middlesbrough[,$_{ESMA}$] had admitted murdering Amanda Jane Fulcher[,$_{SSMA}$] 21[,$_{SSMN}$] a waitress[,$_{ESMA}$] after persuading her to go to the house of a friend[,$_{SSMN}$] Steven Staton[,$_{SSMA}$] 20[, who_{SSCCV}] was jailed for six years for helping dispose of the body.

In this case, the fourth comma, which follows *Middlesbrough*, is the rightmost boundary of the subordinated constituents *Middlesbrough*; *Grangetown, Middlesbrough*; *of Grangetown, Middlesbrough*; and *25, of Grangetown Middlesbrough*. This potential coordinator should be assigned to the class indicating the end of the most superordinate (and largest) of the subordinate constituents, in this case the phrase, *25, of Grangetown Middlesbrough*. It is thus annotated as a case of ESMA, the rightmost boundary of a subordinated adjectival phrase.

8.2 Coincident Rightmost/Leftmost Subordination Boundaries

In Sentence 117, the class of the second markable, *, that* is ambiguous as it serves both as the rightmost boundary of the previously introduced subordinate noun phrase and as a complementiser introducing the clausal complement (i.e. a leftmost boundary). The classification of this markable is made on the basis of the rewriting process. As an obligatory argument of the verb *to tell*, the simplification process should not delete the clause or dislocate it from the main verb. The markable is therefore annotated as the rightmost boundary of the subordinate noun phrase:

(117) Mr Jones told Ronald Bartle, the deputy chief Metropolitan stipendary magistrate[$_{ESMN}$, that] under the European convention on extradition signed by both Britain and Spain, it was not the duty of the court to decide General Pinochet's guilt, nor did Spain have to prove there was a case to answer Mr Bartle simply had to assess whether the general had been accused of an extraditable crime.

This example shows a case of indirect speech that is used to report statements, and takes the form of a nominal *that*-clause.

8.3 Coincident Coordinator/Subordination Boundary

As shown in Section 7.3 (Example 114), the boundaries of conjoins also serve as the boundaries of any subordinate constituents of those conjoins. This is true for all types of conjoin, whether subordinated or coordinated.

8.4 Ill-Assorted Coordination

In general, it is assumed that coordination occurs between conjoins that match in terms of form, function, and meaning. This usually means they are of matching syntactic categories. However, there are a proportion of instances demonstrating "ill-assorted" coordination [Quirk et al., 1985]. Sentence (118) demonstrates ill-assorted coordination of an Adjective Phrase and a Prepositional Phrase.

(118) Whether it was unlawful depended on whether the nursing services were "merely incidental or ancillary to the provision of accommodation which a local authority is under a duty to provide" [and] "of a nature" which a local authority providing social services could be expected to provide.

Sentence (119) demonstrates ill-assorted coordination of an adjective phrase[7] and a verb phrase.

(119) "Name something that is currently on BBC1 that gets people excited [and] talking about it.

It is hypothesised that the automatic simplification of sentences involving ill-assorted coordination will be unreliable. For this reason, these instances should be assigned the class SPECIAL until a suitable methodology for their automatic simplification has been developed. Due to the scarcity of this class, it will be considered outside the scope of 287607 FIRST.

8.5 Syntactic Ambiguity/Uncertainty

The annotation task can be hindered by cases of genuine ambiguity and justifiable uncertainty. For example, in (120), there is uncertainty about the syntactic category of the subordinated constituent. It seems to function as an adverb, but its head is verbal. In (121), it is difficult to derive the relation of the subordinated constituent to the main clause.

(120) "It's a massive thing for anyone to face[,] let alone a teenager - at that age it's very difficult."

(121) Not immediately[, but] it has opened the debate and put pressure on ministers to make clear who can expect to receive free long-term care.

It is recommended that instances demonstrating genuine ambiguity and uncertainty be annotated as belonging to the class SPECIAL.

8.5.1 Verbs/Adjectives

Annotators sometimes classify instances such as the one highlighted in Sentence (122) as coordinated adjective phrases (CMA1) as opposed to verb phrases (CMV1).

Sentence (122) demonstrates coordination of adjectival phrases postmodifying the noun *murder*.

[7]This categorisation of *excited* is supported by the fact that insertion of the modifier *very* prior to the first conjoin generates an acceptable sentence whereas insertion prior to the second does not.

(122) A father of four was jailed for life yesterday for a murder committed 15 years ago[, but] unsolved until DNA tests proved he was the killer.

This can be avoided by applying a test of grammaticality. If insertion of the word *very* before the participle produces an acceptable sentence, then the instance should be considered to be coordinating participial adjectives and be annotated as CMA1. If not, then it should be classified as CMV1.

There is some ambiguity in (123) about whether *fighting* is being used as an adjective, with no direct object, or whether it takes *each other* as a direct object and is being used as a verb. This depends on the referent of the pronoun *they*, either as an entire collective, or as one faction within the collective.

(123) He said: 'They were fighting[,] kicking and punching each other.'

If this ambiguity cannot be resolved by inspection of the source document, annotators may apply a heuristic that the part of speech of *fighting* matches the parts of speech of the other coordinated elements in this multiple coordination.

8.6 Failure to Identify Combinatory Coordination

In a significant proportion of cases, annotators find it difficult to identify instances of combinatory coordination (see examples in Appendix 3.21). To avoid this problem, annotators are encouraged to consider the form of the simplified sentences that will be automatically derived from the original. If the simplified sentences would be inconsistent with the meaning of the original, then the instance in the original sentence should be annotated as COMBINATORY.

8.7 Limitations of the Annotation Scheme

The coordinators in Sentences (124) and (125) were initially annotated as instances of CMV2. However, these sentences motivate the addition of a new class in a revised version of the annotation scheme. This class will be CCV2, the coordination of clauses in which the head of one conjoin (i.e. the head verb) is elided.

(124) Alden was given$_i$ 15 years [and] Wright ϕ_i eight years.

(125) The 38-page judgment stated that Mrs Coughlan, a tetraplegic, was entitled to free nursing care because her primary need for accommodation was$_i$ a health need [and] her nursing needs ϕ_i not 'incidental'.

Currently, these instances should be annotated as SPECIAL given that the existing simplification method for the class CCV is not directly applicable to these cases.

Sentence (126) demonstrates a complex case of VP coordination with elision of both the head verb *see* and the adverbial modifier *inside*.

(126) "Looking through a window they could see$_i$ swarms of flies inside$_j$[, and] ϕ_i a cat in obvious distress ϕ_j.

Again, given that the current simplification algorithm is not able to process such constructions, these instances should be annotated as SPECIAL.

9 Appendix: Classes

The labels used to denote different linking functions are acronyms.

- The first part of the acronym can indicate the rightmost boundary of a subordinated constituent (ES), the leftmost boundary of a subordinated constituent (SS), or a coordinating function (C).
- The second part of the acronym indicates the **syntactic projection level** of the linked conjoins. This may be morphemic (P), lexical (L), maximal, i.e. phrasal, (M), or clausal (C).
- The third part of the acronym indicates the **syntactic category** of coordinated conjoins or the subordinate constituent bounded by the potential coordinator. This may be verbal (V), nominal (N), adjectival (A), adverbial (Adv), prepositional (P), or Quantificational (Q).
- The fourth part of the acronym is used to distinguish between classes that cannot be distinguished on the basis of the previously mentioned grounds. For instance, to distinguish between verb phrase coordination in which the head verb has been elided from those in which no such elision occurs.
- Additional symbols are used to annotate the left and right boundaries of interjections (SSMI/ESMI), direct quotes (SSCM/ESCM), punctuation introducing tag questions (STQ), and the linking of elements that cannot be split for the purpose of the syntactic simplification process (COMBINATORY). Instances that are otherwise unclassifiable are annotated as SPECIAL. When an annotator lacks confidence in assigning a class to any particular instance, s/he can use the label HELP. This class is included in order to save time and to enable batch review of those instances with other annotators/experts.

References

[Quirk et al., 1985] Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.