Corpus linguistics and signed languages: no lemmata, no corpus

Trevor Johnston
Department of Linguistics, Macquarie University
Sydney, New South Wales, Australia
E-mail: trevor.johnston@mq.edu.au

Abstract

A fundamental problem in the creation of signed language corpora is lemmatisation. Lemmatisation—the classification or identification of related word forms under a single label or lemma (the equivalent of headwords or headsigns in a dictionary)—is central to the process of corpus creation. The reason is that signed language corpora—as with all modern linguistic corpora—need to be machine-readable and this means that sign annotations should not only be informed by linguistic theory but also that tags appended to these annotations should be used consistently and systematically. In addition, a corpus must also be well documented (i.e., with accurate and relevant metadata) and representative of the language community (i.e., of relevant registers and sociolinguistic). All this requires dedicated technology (e.g., ELAN), standards and protocols (e.g., IMDI metadata descriptors), and transparent and agreed grammatical tags (e.g., grammatical class labels). However, it also requires the identification of lemmata and this presupposes the unique identification of sign forms. In other words, a successful corpus project presupposes the availability of a reference dictionary or lexical database to facilitate lemma identification and consistency in lemmatisation. Without lemmatisation a collection of recordings with various related appended annotation files will not be able to be used as a true linguistic corpus as the counting, sorting, tagging. etc. of types and tokens is rendered virtually impossible. This presentation draws on the Australian experience of corpus creation to show how a dictionary in the form of a computerized lexical database needs to be created and integrated into any signed language corpus project. Plans for the creation of new signed language corpora will be seriously flawed if they do not take this into account.

1. Introduction

After a brief discussion of the nature and role of corpora in contemporary empirical linguistics, I describe the Auslan (Australian Sign Language) Corpus and the Auslan Lexical Database. I discuss what makes this a genuine linguistic corpus in the modern sense: lemmatisation (Kennedy, 1998). Lemmatisation of signs in the corpus is made possible by the existence of the Auslan Lexical Database. It is an indispensable aid to consistent sign identification through glossing. Lexical information found in the Auslan Lexical Database is being integrated into the annotations of the corpus of Auslan texts. I follow the discussion of the corpus and database by describing some of the annotation conventions observed in the Auslan Corpus that allow for the lemmatisation of lexical signs and, equally importantly, the conventions observed in the annotation non-lexical signs. Together both sets of practices and conventions ensure that the corpus becomes, and remains, machine-readable as it is enriched over time.

2. Corpora and empirical linguistics

Signed language corpora will vastly improve peer review of descriptions of signed languages and make possible, for the first time, a corpus-based approach to signed language analysis. Corpora are important for the testing of language hypotheses in all language research at all levels, from phonology through to discourse (Baker, 2006; McEnery *et al*, 2006; Sampson, 2004; Sinclair 1991). This is especially true of deaf signing communities which are also inevitably young minority language communities. Although introspection and observation can help develop hypotheses regarding language use and structure, because signed languages lack written forms and well developed community-wide standards, and have interrupted transmission and few native speakers, intuitions and researcher observations may fail in the absence of clear native signer consensus of phonological or grammatical typicality, markedness or acceptability. The past reliance on the intuitions of very few informants and isolated textual examples (which have remained essentially inaccessible to peer review) has been problematic in the field. Research into signed languages has grown dramatically over the past three to four decades but progress in the field has been hindered by the resulting obstacles to data sharing and processing.

Moreover, as with all modern linguistic corpora, it should go without saying that signed language corpora should be representative, well-documented (i.e., with relevant metadata) and machine-readable (i.e., able to be annotated and tagged consistently and systematically) (McEnery & Wilson, 1996; Teubert & Cermáková, 2007). This require dedicated technology (e.g., ELAN), standards and protocols (e.g., IMDI metadata descriptors), and transparent and agreed grammatical tags (e.g., grammatical class labels) (Crasborn *et al*, 2007). However, it also requires the identification of lemmata. Lemmatisation—the classification or identification of related forms under a single label or lemma (the equivalent of headwords or headsigns in a dictionary)—is absolutely fundamental to the process of corpus creation. A successful corpus project team should already have available a reference dictionary or lexical database to facilitate lemma identification and consistency in lemmatisation. Without lemmatisation a collection of recordings (digital or otherwise) with various related annotation files (appended and integrated into a single multimedia file or simply related to each other in a database) will not be able to be used as a true linguistic corpus as the counting, sorting, tagging. etc. of types and tokens is rendered virtually impossible.

Annotations began in 2005 and it is anticipated that it will take at least 10 years for a substantial number of these texts to be sufficiently richly annotated for extensive corpus-based research. However, given that corpus-based signed language studies are beginning from such a low base (essentially zero), a recent initial study of 50 annotated Auslan texts from this corpus is already one of the largest of its kind (Johnston *et al*, 2007). A second corpus-based study on the co-occurrence of pointing signs with indicating verbs is being presented at this conference (de Beuzeville & Johnston, this volume).

3. The Auslan Corpus

The corpus brings together into one digital archive a representative sample of a signed language in which the video recordings themselves, along with appended metadata and annotation files, are openly accessible.[1] Importantly, the annotation files of the corpus are designed to facilitate expansion and enrichment over time by various researchers through repeated annotation parses of individual texts.

The Auslan Corpus is built from two sources: the Sociolinguistic Variation in Auslan Project (SVIAP)[2] and from the Endangered Languages Documentation Project (ELDP)[3]. Both datasets are based on language recording sessions conducted with native or near-native users of Auslan. The SVIAP corpus consists of films of 211 participants from the five major cities in Australia (Sydney, Melbourne, Brisbane, Adelaide and Perth). This yielded over 140 hours of unedited digital video footage of free conversation, structured interviews, and lexical sign elicitation tasks. The ELDP yielded approximately 300 hours of unedited footage taken from 100 participants from the same five cities. Each participant was involved in three hours of language-based activity that involved an interview, the production of narratives, responses to survey questions, free conversation, and other elicited linguistic responses to various stimuli such as a picture-book story, a filmed cartoon, and a filmed story told in Auslan. This footage has been edited down to around 150 hours of usable language production which, in turn, has been edited into approximately 1,700 separate digital movie texts for annotation. To date approximately 100 of these texts have been annotated using ELAN (EUDICO Linguistic Annotator) (Hellwig et al., 2007). In total, the corpus consists of digital movies, ELAN annotation files and IMDI metadata files (Johnston & Schembri, 2006).

4. The Auslan Lexical Database

The Auslan Lexical Database, consists of over 7,000 individual sign entries and was originally created as a FileMaker Pro database file (Johnston, 2001). Lexical signs in the form of short digital movie clips are headwords/lemmas of individual records/entries in the database. There are multiple fields coding information on the form, meaning and lexical status of each headsign. Form fields include one for phonological transcription using modified HamNoSys, several for dedicated feature fields coding for handshape, location, symmetry, etc.; and one field for morphological transcription which relates variants to stem forms. Meaning fields include several for definitions, semantic domains, and synonyms and antonyms. Lexical status fields include several for dialect, register, and stem/variant identification. The database lists a citation form of a lexical sign as a major stem entry, with common variant forms listed separately.

This database also now exists in two other forms: (i) an online, open access dictionary called *Auslan Signbank* (http://www.auslan.org.au) and (ii) a limited access researchers' reference database which also includes variant signs and newly identified signs. The database, in both its current forms, is being constantly corrected and augmented. Finally, signs in the database are organized and sequenced formationally, i.e., according to major phonological features of signs, such as handshape and location, so that scrolling through the database records displays formationally similar signs one after the other.

The Auslan Lexical Database is the source of information for a number of dictionaries of Auslan in three formats—print, CD-ROM, and internet (e.g., *Auslan Signbank*, mentioned above). By definition, the sign data is lemmatised. It serves as the reference point for the lemmatisation of the corpus annotations. However, since the identification of lexis in any language is always open-ended, it should be noted that corpus data is also used to test assumptions underlying the lemmatisation found in the Auslan Lexical Database itself. In other words, the source database and annotations are appropriately updated as required (as described below). This strategy is one possible solution to the 'database paradox' (van der Hulst *et al*, 1998).

5. Lemmatisation in the Auslan Corpus

In order for a corpus of recordings of face-to-face language in either spoken or signed modalities to be machine readable, time-aligned annotations need to be appended to the source data using some form of multi-media annotation software. It is these appended annotations which are read by machine, not the source data itself. Strictly speaking, therefore, a written transcription of the text need not be created in order to do corpus-based research. However, just as with the Auslan Lexical Database, such a level of representation would be necessary in order to

[1] Open-accessibility will be implemented after an initial limited access period of three years from the time of the deposit of the corpus at SOAS in 2008.

[2] Australian Research Council research grant awarded to Adam Schembri and Trevor Johnston — #LP0346973 *Sociolinguistic Variation in Auslan: Theoretical and applied dimensions.*

[3] Hans Rausing Endangered Languages Documentation Program (School of Oriental and African Studies, University of London) language documentation project awarded to Trevor Johnston — #MDP0088.

carry out phonetic or phonological research of a corpus.

With respect to identified sign units, failure to integrate lexical information into the sign identifier, either as a transcription or a gloss-based annotation, immediately creates two problems: (1) the consistency and commensurability of data that is transcribed or glossed by multiple researchers or even the same researcher on different occasions; and (2) the effective unboundedness of the sign dataset. In other words, each sign articulation which may be distinctive would have its own distinctive transcription because each form would have its own representation, or its own distinctive gloss reflecting contextual meaning. The unique identification of sign types—lemmas—would thus not been achieved and one of the prime motivations for the creation of a linguistic corpus in the modern sense would be undermined from the very outset.

5.1 ID-gloss vs. GLOSS vs. translation

Lexical signs need to be identified using a gloss which is intended to uniquely identify a sign. In the Auslan Corpus project this is referred to as the *ID-gloss*. An ID-gloss is the (English) word that is used to label a sign all of the time within the corpus, regardless of what a particular sign may mean in a particular context or whether it has been systematically modified in that context. For example, if a person signs HOUSE (a sign iconically related to the shape of a roof) but actually means *home*, or performs a particularly large and exaggerated form of the sign HOUSE, implying *mansion*, (without that modified form itself being a recognized and distinctive lexeme of the language) then the ID-gloss *house* would still be used in both instances to identify the sign in the annotation.

A consistently applied label of this type means it is possible to search through many different ELAN annotation files and find all instances of a sign to see how and when it is used. Only if a sign always has the same ID-gloss can we search, using computers, for how that sign is used in different ways in the corpus.

The ID-gloss is thus not meant to be a translation of meaning. So if the signer produces SUCCESS but means ‘achieve something’, it is still annotated with the ID-gloss SUCCESS; and if a person signs IMPORTANT but means ‘main’ or ‘importance’, it is still labeled IMPORTANT.

This is crucial. Without consistency in using the ID-gloss it will be impossible to use the corpus productively and much of the time spent on annotation will be effectively wasted because the corpus will cease to be, or never become, machine readable in any meaningful sense. It will not actually be the type of corpus that linguists want to have access to, i.e., a machine readable set of annotated and linguistically tagged texts (which are also representative samples of a language). It will just be a collection of reference texts, a corpus in the ‘old fashioned’ sense.

With respect to distinguishing between glossing and translation, meaning is assigned to the text through glossing only indirectly through the unavoidable fact that the ID-gloss, which is primarily intended to identify a sign, actually uses an English word (or words) that bears a relationship to meaning of the sign. In other words, the ID-gloss is not chosen arbitrarily or randomly. It is highly motivated. However, it is not intended as a translation because within the ELAN annotation files of the corpus, translations are made on their own dedicated tiers. In assigning an ID-gloss we are simply labeling a sign so that it can be uniquely and quickly identified for subsequent tagging with linguistic markers (e.g., for grammatical class, sign modification potential, presence or absence of constructed action, semantic roles, and so on) during a later annotation parse, or searched for with or without these tags being taken into consideration. Apart from the obvious motivation of the English word used to gloss a sign, no serious attempt is being made in the assigning of an ID-gloss to translate a sign.

5.2 Selecting the appropriate ID-gloss for a sign

Annotators refer to the dictionary of Auslan in one of two forms—Auslan Signbank (www.auslan.org.au) or the Auslan Lexical Database (a FileMaker file)—to view signs and their assigned annotation ID-gloss.

If a sign in the text being annotated appears to be a lexical sign and cannot be not found in the dictionary, the annotator chooses a simple English word to gloss that sign as appears to be appropriate. If the annotator cannot avoid using a word that has already been used in the dictionary as an ID-gloss they append a distinguishing number after the gloss. Thus, if HOUSE already exists in the dictionary as the ID-gloss of a sign (and there is also no ID-gloss currently used that is HOUSE2) then the new ID-gloss would be HOUSE2. Similarly, if HOUSE2 already existed as an ID-gloss, HOUSE3 would be created. After an annotation parse has been completed and the ELAN annotation file is submitted back to the corpus managers, the dictionary is updated, if necessary. For example, if a new sign is recognized as a new unrecorded sign, a new dictionary entry will be created with its own distinct ID-gloss (which may or may not be the same as the ID-gloss suggested by the original annotator).

The only time an existing sign form will be assigned a different ID-gloss is when corpus data justifies the identification of a completely distinct and unrelated meaning for the sign form. In such cases the sign form receives its own distinctive the ID-gloss and the two signs are treated as homonyms.

5.3 Annotation conventions: ID-glosses

The consistent use of the same ID-gloss for the same sign is the single most important act in building a machine-readable sign language corpus. It is reinforced by the adherence to a relatively small set of annotation and glossing conventions that ensure that similar types of signs are glossed in similar ways. The following are just a few indicative examples of these types of conventions.

Negative incorporation If a sign incorporates a negative as part of its meaning, the main verb gloss is given first followed by a gloss for the negative element. This makes it easier to search and sort signs by meaning and name (e.g., KNOW and KNOW-NOT will be next to each other if sorted alphabetically or both will be found if sub-string search routines are used). It also means all negative incorporation is expressed the same way, rather than sometimes with words like DON'T (e.g., if glossed as DON'T-KNOW rather than KNOW-NOT) or sometimes with an entirely different word form, such as WON'T for WILL-NOT.

Variant forms Sometimes a sign form is clearly recognizable as a minor variant of a more common or standard form, using a slightly different handshape, movement pattern or location. These minor variations are not normally reflected in any change to the ID-gloss. Generally speaking, one does not want there to be an unnecessary proliferation of ID-glosses through attempts to encode in the gloss itself information about formational variation. Many of the possible variant forms of many signs have already been recorded in Auslan Lexical Database and are well understood. Therefore, the ID-gloss assigned these variant forms is often the same as the citation or unmarked form. However, if phonetic or phonological analysis is the focus of the annotations being created, specific phonological tiers in the ELAN annotation templates can be used utilized for this purpose. On these tiers, transcription using dedicated fonts such as HamNoSys can be used to capture the actual form of the sign. Alternatively, if the variant form noted in the textual example is unrecorded in the Auslan Lexical Database and appears to be particularly noteworthy and is not part of some grammatical modification that will be recorded on other tiers of the annotation, a brief addition to the ID-gloss can encode this. In these cases, a letter code of the handshape change is added after a hyphen (e.g., SUGAR-K would signify the sign SUGAR made with a K handshape), or a word for the variant location or the variant movement is addd (e.g., KNOW-cheek signifies KNOW made on the cheek). However, all such additions to any ID-gloss should be kept to an absolute minimum and should not be done in a way that would confound search and sorting routines.

Numbers If a signer uses a number to refer to anything it is annotated using wordS, not digits. For example, *NINETEEN-EIGHTY-SEVEN* rather than 1987, *FOURTEEN-YEARS-OLD* rather than 14-years-old.

Points All ID-glosses for points begin with the initials *PT* (for 'point'). This allows for all pointing signs in the corpus to be identified regardless of the grammatical function that may or may not be attributed to them by various annotators. Indeed, this glossing convention enables one to collect and compare all instances of points, facilitating their subsequent relabelling if textual evidence justifies reanalysis. Further grammatical details are given whenever possible (e.g., *PT:PRO* signifies 'pointing sign functioning as a pronoun', *PT:DEM* signifies 'pointing sign functioning a demonstrative pronoun', and *PT:POSS* signifies 'pointing sign functioning a possessive pronoun'). Indeed, annotations may be even more detailed. For example, *PT:PRO3pl* signifies 'pointing sign as a third person plural pronoun'. If the handshape changes from what is normally expected, that information is included immediately after the pt, in parentheses. For example, *PT(B):PRO1sg* signifies 'first person singular made with a flat handshape'. However, in many cases, it will be difficult, or even impossible, for an annotator to be able to make a very detailed grammatically rich annotation with certainty. Provided the convention of ID-glosses for pointing signs beginning with *PT* is adhered to then decisions about the actual function of certain pointing signs can be deferred until more textual examples are collected.

Sign names Sign names are prefixed with *sn:* followed by the proper name in lower case. Thus a sign name for a person called Peter would be written as *sn:peter*. Additional information may be added, but is not required. For example, if the sign name is based on fingerspelling the relevant letter(s) and a hit regarding sign form can be added after the gloss, thus: *sn:peter(-P-shake).* If the sign name is identical in form to a lexical sign the relevant sign may be identified after the name in brackets: *sn:peter(ROCK).*

Foreign borrowings Lexical signs which are clearly recent or idiosyncratic borrowings from another signed language and which are generally not considered to be Auslan signs are given best gloss possible followed by the name of the signed language. For example, the borrowed sign *COOL* from ASL would be written as *COOL(ASL)*

5.4 Lexical vs. non-lexical signs

Lemmatisation can only apply to lexical signs. However, may signed meaning units found in natural signed language texts are not lexical signs. As a number of signed language linguists have noted one needs to distinguish at two major types of meaning units—lexical signs and non-lexical signs (e.g., Johnston & Schembri, 1999; Sandler & Lillo-Martin, 2006). *Lexical sign* is reserved for a form whose meaning in context is more than the conventionalized and/or iconic value of its components (handshape, location, etc.) within the inventory of meaning units of a given signed language in a given context, and that meaning is consistent across contexts. It is essentially, equivalent to the commonsense notion of *word* (Sandler & Lillo-Martin, 2006). The term *non-lexical sign* is reserved for a form that has little or no conventionalized or language-specific meaning value beyond that of its components in a given context (e.g., depicting or 'classifier' signs).

5.4.1 Annotation conventions: non-lexical signs

As with ID-glosses, a relatively small set of annotation and glossing conventions need to be adhered to in order to ensure that similar types of non-lexical signs are glossed in similar ways. Without such conventions, these categories of signs cannot be easily extracted from the corpus for analysis and comparison. The following are just a few indicative examples of these types of conventions.

Depicting signs These 'do it yourself' signs are not listed in signed language dictionaries because their meaning is too general or context specific to be given a meaningful entry description. In the Auslan corpus all such signs begin with *pm* (for "property marker") as the handshape shows a property of the object.[4] Since handshape is a very salient feature of depicting signs it is included in the annotation gloss for these types of signs in the following format — *pm(handshape):brief-description-of-meaning-of-sign*. For example an upright index finger representing the displacement of a person would be annotated thus: *pm(1):person-walks*. One does not need to annotate full details of the form of the depicting sign in order to create a grammatically useful annotation because the form of the sign is visible in the video that is always attached to the ELAN annotation file. However, should such information be important, it belongs on separate tiers of the annotation file dedicated to encoding phonetic and phonological information about individual signs.

List buoys A list buoy is a hand which is held throughout a stretch of discourse, usually on one's left (or weak) hand, and uses count handshapes to mark the movement to each of a sequentially related set of entities or ideas. The handshape can be held in space throughout the articulation of each item, or appear and reappear if two-handed signing demands it be removed in order to produce certain signs. The signer usually grabs or points to a relevant finger of the buoy for each item in the list. The buoy is prefixed with *buoy* (or simply the letter *b* for 'buoy') followed by a label of the handshape being used in brackets and, after a colon, a short description of what it stands for. So an index finger held up to indicate the first of a series of items would be annotated: *buoy(1):first-of-one* or *b(1):first-of-one*. As each finger is added for each item they are annotated accordingly in turn: *buoy(2):second-of-two* or *buoy(3):third-of-three*. If the handshape anticipates all of the members of a series by holding up two, three, four, or five extended fingers throughout, the range is stated: *buoy(8):three*. In this latter case especially, but it is also possible in the other instances, the dominant hand may simultaneously point at a specific finger of the buoy (or it may hold it). This is annotated on the dominant hand according to the finger identified and whether it is a pointing or holding action (e.g., *PT:buoy-third-of-five* or *HOLD:buoy-third-of-five)*. If the dominant simply points to the entire buoy, it is annotated as *PT:buoy*. There is no need to repeat information about the buoy itself (handshape and/or number of entities) on the annotation for the dominant pointing hand because the annotation for the subordinate (weak) hand will have that information about the buoy already coded.

Fingerspelling Any time a signer uses fingerspelling, the word is prefixed with *fs:* for 'fingerspelling' followed by the word spelled, thus— *fs:word*. If not all the letters of a word are spelled, and it is clear what that word is, the omitted letters are put in brackets—*fs:wor(d)* not *fs:wor.* If the fingerspelling is for multiple words, a new annotation is begun for each word even if it is one continuous act of fingerspelling—*fs:mrs fs:smith* not *fs:mrssmith*. By following these conventions, it is easier for the number of fingerspellings to be counted and the types of words that are fingerspelled to be identified. If the form of a lexical sign is a single fingerspelled letter which could mean various things, the letter is followed by the word it stands for— *fs:m-month*, *fs:m-minute*, *fs:m-mile*.

5.4.2 Annotation conventions: gesture

A gesture is neither a lexical sign nor a non-lexical sign. Gestures are quite common in naturalistic signing. As with depiciting signs, when identifying or glossing a gesture one need not describe the form of the gesture on the sign identification (glossing) tier. The form of the gesture is visible in the associated movie or can be coded on separate dedicated phonetic or phonological tiers in the annotation file. One would thus write something like *g:how-stupid-of-me* not *g:hit-palm-on-forehead.*

6. Conclusion

No claim is being made here that the specific glossing conventions used in the Auslan Corpus should form the basis of a standard for all signed language corpora. Though consistency across signed language corpora in annotation protocols would facilitate cross linguistic comparisons and thus be extremely desireable, the most important considerations in the first instance are the principles of lemmatisation and consistent treatment (glossing) of various sign types.

However, there is no escaping the observation that any attempt to build a linguistic corpus, in the modern sense, of a signed language without reference to, or without the prior existence of, a relatively comprehensive lexical database of the language in question could well be plagued by difficulties. It would be extremely difficult, if not impossible, to control the proliferation of glosses referring to the same sign without a lexical database that is arranged by, or searchable on, formational or phonological criteria. This principle is fundamental to the entire enterprise of corpus creation in signed language linguistics. Without lexical resources of this type, plans to create signed language corpora are unlikely to produce anything resembling what is today commonly understood by a linguistic corpus.

[4] This terminology is borrowed from Slobin and Hoiting. However, any abbreviation, consistently applied, would be appropriate (e.g., *cl:* for 'classifier sign', or *d:* for 'depicting sign').

Linguists need to be able to identify each sign form uniquely and this must be done by sorting sign forms phonologically. This is the role of the lexical database. Without this, one could not locate and compare sign forms in order to determine if a new unique gloss is required for a particular sign form rather than just the association of an additional sense to an existing one. Once again this is a piece of information to be added the lexical database, not included in the annotation at the ID-gloss level. To a computer using searching or sorting routines on a corpus, non-uniquely identifying glosses would be next to useless.

The lexical database and its representation in dictionaries in various forms, is thus an unavoidable prerequisite for creation of a viable corpus. However, it need not be exhaustive. After all, it is highly likely a corpus will actually reveal unrecorded lexical signs which need to be added to the reference lexical database.

7. References

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., van der Kooji, E., Woll, B., et al. (2007). Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics, 12*(4), 535-562.

Hellwig, B., van Uytvanck, D., & Hulsbosch, M. (2007). EUDICO Linguistic Annotator (ELAN). http://www.lat-mpi.eu/tools/elan/

Johnston, T. (2001). The lexical database of Auslan (Australian Sign Language). *Sign Language & Linguistics, 4*(1/2), 145-169.

Johnston, T., & Schembri, A. (1999). On defining lexeme in a sign language. *Sign Language & Linguistics, 2*(1), 115-185.

Johnston, T., & Schembri, A. (2006). Issues in the creation of a digital archive of a signed language. In L. Barwick & N. Thieberger (Eds.), *Sustainable data from digital fieldwork: Proceedings of the conference held at the University of Sydney, 4-6 December 2006* (pp. 7-16). Sydney: Sydney University Press.

Johnston, T., & Schembri, A. (2006). *The use of ELAN annotation software in the Auslan Archive/Corpus Project.* Paper presented at the Ethnographic Eresearch Annotation Conference, University of Melbourne, Victoria, Australia (Feburary 15-16).

Johnston, T., de Beuzeville, L., Schembri, A., & Goswell, D. (2007). *On not missing the point: Indicating verbs in Auslan.* Paper presented at the 10th International Cognitive Linguistics Conference, Kraków, Poland (15-20 July).

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London and New York: Longman.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (Eds.). (2006). *Corpus-Based Language Studies*. London and New York: Routledge.

Sampson, G. (2004). *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Teubert, W., & Cermáková, A. (2007). *Corpus Linguistics: A Short Introduction*. London: Continuum.

van der Hulst, H., Crasborn, O., & van der Kooij, E. (1998, December). *How SignPhon addresses the database paradox.* Paper presented at the Second Intersign Workshop, Leiden, The Netherlands.