Adding value to, and extracting of value from, a signed language corpus through secondary processing: implications for annotation schemas and corpus creation

Trevor Johnston
Macquarie University
Sydney, Australia
E-mail: trevor.johnston@mq.edu.au

Abstract

A basic signed language (SL) corpus is created through primary processing of video recordings using multi-media annotation software. Primary processing entails the tokenization and identification of SL units. For the purposes of linguistic research a corpus also needs secondary processing. Secondary processing entails appending tags for specific linguistic features to primary annotations. I draw on the experience from the Auslan corpus project to describe how primary and secondary processing can be used in corpus-based SL research. In particular, I show how the tier structure of ELAN can be used to tag SL units in a variety of ways, and how this information can be used to glean new information from the corpus which can then be added as new annotations to the corpus. Value-adding by principled and systematic primary and secondary processing of digital recordings is thus not only essential for corpus creation ('machine-readability'), it also enables further enriching of the corpus so that even more value can be extracted. I conclude by discussing the implications for annotation software and standardized annotation schemas used in the creation of SL corpora.

The case for SL corpus linguistics

There are many arguments that have long been advanced in support of corpus-based language description and linguistic research and they all apply equally well to SLs. There is no time to repeat them here even if they do go to the very heart of what it is linguists treat as (sufficient) evidence for a claim about the grammar of a language. Suffice it to say, however, that I take them as strong arguments in favor of basing descriptive and theoretical linguistics on how people use a particular language, and not on their intuitions or judgments (at least, not alone). However, there are several additional reasons why corpora are particularly important in SL research, and some of them are unique to this field of linguistics. They do bear repeating, e.g. see Johnston & Schembri (2010).

SLs are languages of minority communities that rarely have any real geographical centre, apart from perhaps residential schools for the deaf or deaf clubs. SLs experience interrupted inter-generational transmission for all but a tiny minority of users and thus have few native users. SLs have no dedicated or widely used written form, nor long history of being used in education.

These facts create two major problems for SL researchers. First, intuitions may be less useful in language description work in SL-using communities, all of which have been characterized as displaying high degrees of variation in both lexis and grammar. Moreover, users sometimes appear to lack sets of shared linguistic norms that are often found in stable language communities, especially those with literacy and standard varieties used in education. This variability means there may be little consensus on phonological or grammatical typicality, markedness or acceptability among users. The practice in SL linguistics of relying on the intuitions of a small number of informants can thus be seen as problematic (even if one was to give high evidential status to intuitions and/or grammaticality judgments in the first instance). Second, the representation of SL utterances using written glosses has meant that primary data have remained essentially inaccessible to other researchers and consequently unavailable for meaningful peer review.

In short, there is a particular need for SL recordings which can be processed into language corpora in order to empirically ground our understanding of the structure, use, acquisition and learning of SLs, and to test claims or hypotheses about their grammars. Without corpora, one risks basing educational interventions and interpreting training, the design of automatic SL processing or recognition systems, and even linguistic theory itself on descriptions of SLs that may be inadequate.

Language processing and corpus linguistics

In the history of SL research almost no extended SL texts of any kind have been created, either by glossing or by using a dedicated notation system, that could in turn be digitized, read by computer and further processed.

With recent advances in digital recording technology, computing, and multimedia annotation software, the way in which recordings of face-to-face language could be best processed to create corpora for the purposes of linguistic analysis has been transformed (cf., Beal, Corrigan & Moisl, 2007). For instance, the source text can now remain the primary data itself, rather than being necessarily replaced by its representation in a transcription to which annotations were subsequently appended. This has made the creation of SL corpora feasible. One of a number of multimedia annotation software programs suitable for use by SL researchers wishing to create corpora is called ELAN (Max Planck Institute for Psycholinguistics Language Archiving Technology Group, 2009).

A minimalist corpus: primary processing

A basic signed language reference corpus is created through primary processing of the raw video recordings in an archive using multi-media annotation software, e.g. ELAN. Primary processing entails the tokenization and identification of signed units. This can be achieved by ensuring that conventional linguistic units and types are systematically and consistently identified with invariant and unique sign identifiers (or "IDglosses"). Consistency in type/token identification is the key requirement for ensuring that a SL corpus is machine-readable for the purposes of linguistic research (see Johnston, 2010).

This is achieved by corpus annotators adhering to set protocols and schemas with respect to the classification and identification of sign types and the assignment of IDglosses to fully lexical signs. The Auslan corpus project has developed such a set of guidelines and other SL corpus projects are in the process of developing their own.[1] In SL corpora, attention must be paid to distinguishing between fully-lexical signs and partly-lexical signs (both content signs and grammatical signs) and gestures (both manual and non-manual).

A minimalist corpus also usually involves the addition of a time aligned parallel translation into the working majority spoken language. Indeed, in some very basic corpora the only annotation may be a parallel translation, grossly time-aligned to the source media.

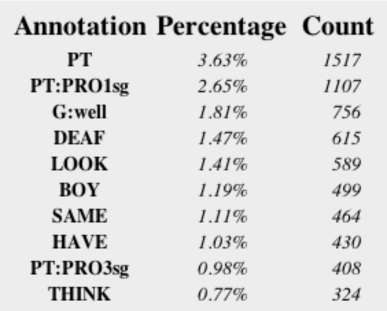
Just on the basis of primary processing of a corpus, it is possible to glean valuable information on sign tokens, sign types, or signs by IDgloss, e.g. number, frequency, duration, and concordance/collocation patterns. It is even possible to conduct preliminary and tentative grammatical analyses, by locating segments of the primary text that co-occur with particular constructions in the translated parallel text.

Before turning to secondary processes, I will briefly exemplify how these primary annotations can be used to extract this type of information in the ELAN search routines. However, partly because of space constraints in this paper and time constraints in the presentation, I will only be able to discuss frequency and collocation.

IDgloss frequency

Selecting within the ELAN menu thus: > Search > Single Layer Search, one defines the search domain (keeping separate left hand dominant from right hand dominant signers), selects the mode (annotation, regular expression), selects the tier (IDgloss) and specifies the search. In this case, **.+** for "any text". There are 41,842 hits in the result of which approximately 10% are represented in the top 10 most frequent IDglosses (Figure 1).

Substring match searches can be used to specify the beginning of an annotation string (such as ^PT "begins with PT", ^DS "begins with DS" and so on). In this way, one can exploit the glossing conventions for partly-lexical and non-lexical signs and gestures to search for these types of signs by general type (e.g., ^PT or "a pointing sign") or more specifically (e.g. ^PT:PRO1SG(7) or "first person singular pointing sign made with and index finger and extended thumb").



Annotation	Percentage	Count
PT	3.63%	1517
PT:PRO1sg	2.65%	1107
G:well	1.81%	756
DEAF	1.47%	615
LOOK	1.41%	589
BOY	1.19%	499
SAME	1.11%	464
HAVE	1.03%	430
PT:PRO3sg	0.98%	408
THINK	0.77%	324

Figure 1: Frequency view of IDgloss search results[2]

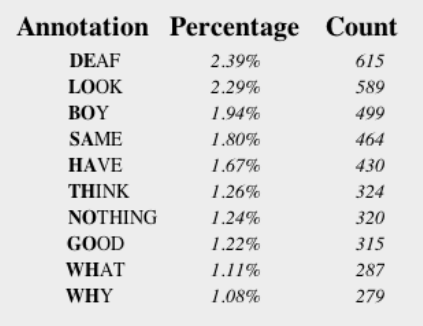
Using this method, it was established by searching the first 'minimalist' annotated texts in the Auslan corpus dating from 2006-07 that approximately 11% of all signs in the corpus were points, 7% were gestures, and 10% were depicting signs (i.e. up to almost 30% of all signs produced were either non-lexical or partly-lexical signs). Interestingly, as the corpus has grown, from 10,000 to 60,000 sign tokens, these relative proportions have changed little.[3]

IDgloss (fully-lexical signs only) frequency

Using the same procedure as in the previous search but with the search text specified as:

^.[^\QPT\E|^\QDS\E|^\QFS\E|^\QG:\E]

for "begins with any text except PT (point), DS (depicting sign), FS (fingerspelling), or G (gesture)" (in other words, "find all lexical IDglosses.") yields all lexical signs. There are 25,750 hits in the result, but only the top 10 are displayed in Figure 2.



Annotation	Percentage	Count
DEAF	2.39%	615
LOOK	2.29%	589
BOY	1.94%	499
SAME	1.80%	464
HAVE	1.67%	430
THINK	1.26%	324
NOTHING	1.24%	320
GOOD	1.22%	315
WHAT	1.11%	287
WHY	1.08%	279

Figure 2: Frequency view of lexical sign search hits

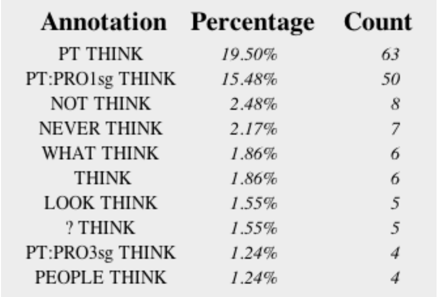
Collocation and frequency

Using the same procedure as in the previous search but with search type specified as *n-gram over annotations* and the search text as # *think* (for "any two sequential annotations, the second of which is THINK"), the results (out of 330 hits) are displayed in Figure 3. (Once again, the table only displays the top 10 hits.)

[1] The Auslan annotation guidelines can be downloaded from http://www.auslan.org.au/about/annotations/

[2] Signs glossed simply as PT have yet to be further specified.

[3] The aim is to expand the corpus to 100,000 sign tokens by the end of 2010 and to double that number again by 2012 by increasing the number of annotated digital movies from the current 201 clip to around 500. There are more than 1,200 movie files in the corpus.



Annotation	Percentage	Count
PT THINK	*19.50%*	*63*
PT:PRO1sg THINK	*15.48%*	*50*
NOT THINK	*2.48%*	*8*
NEVER THINK	*2.17%*	*7*
WHAT THINK	*1.86%*	*6*
THINK	*1.86%*	*6*
LOOK THINK	*1.55%*	*5*
? THINK	*1.55%*	*5*
PT:PRO3sg THINK	*1.24%*	*4*
PEOPLE THINK	*1.24%*	*4*

Figure 3: Frequency view of signs preceding THINK[4]

These searchers are only possible because of distinctions made in the IDglossing between type and token, and between sign sub-types. However, the real efficacy of this type of annotation schema becomes best seen if we look at its place in secondary processing.

A value-added corpus: secondary processing

For the purposes of conducting detailed linguistic research a corpus also needs to undergo secondary processing.

Secondary processing entails appending information to annotations created in primary processing. These secondary annotations (or 'tags') add specific phonological, morphological, syntactic, semantic, pragmatic or discourse information about linguistic forms, depending on the purpose of the analysis. In ELAN the tags are distributed over multiple tiers, each dedicated to a certain type of tag. Once again protocols and schemas need to be implemented to ensure that the tags used are drawn from a limited or controlled vocabulary of values and that they are applied to the primary annotations in a consistent manner. These too are covered in the annotation guidelines for the Auslan corpus.

Secondary processing enables one to extract far more sophisticated frequency statistics for any annotation (IDgloss or linguistic tag) and to specify and identify the environments in which they occur in greater detail. For example, ELAN searches can be constrained by specifying aligned or overlapping values on as many as two other tiers for any specified annotation or string of up to three annotation values. In addition, multiple annotation files can be specified as the search domain. These can be selected manually or automatically based on metadata values such as age, gender, region, text type, etc.

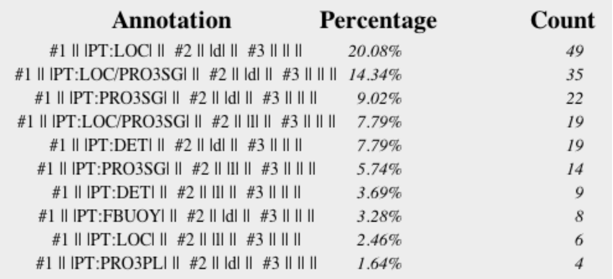
The analysis of the search results can be partially done though examining ELAN's search results directly or by exporting them in various formats. For example, once the matches have been computed they are displayed in either concordance or in frequency views in the ELAN search dialogue box. Both of these data types can then be exported for further processing in various databases or corpus analysis software programs.

With respect to the Auslan corpus, a number of studies are now underway using texts that have been enriched with secondary annotations, be they formational (palm orientation, handshape, sign location and/or sign directionality), lexico-grammatical (grammatical class, argument structure, semantic roles, 'PRO-drop'), and 'utterance' level (clause boundaries, constructed action).

I now describe the procedure that makes it possible to use secondary annotations in the ELAN search routines to extract interesting and relevant linguistic observations. Once again, due to space and time constraints, I give only a few examples—palm orientation, grammatical class, and clause argument structure—as well as briefly discussing constructed action. I only give example data drawn from subsets of the Auslan corpus. A formal report using corpus-wide and definitive data is not my purpose here.

Palm orientation and pointing signs

Selecting from the ELAN menu Search > Multiple Layer Search, one then defines the search domain, selects the mode and the search tiers (1 IDgloss, 2 orientation), and then specifies the search text: ^PT ("begins with PT") for the IDgloss and .+ or "any text" for the orientations (d = down, l = left, u = up, r = right, o = other), as well as specifying that both annotations *overlap*. The results in an example subset of 19 eafs have 244 hits (only top 10 displayed, see Figure 4).



<table>
<tr><th>Annotation</th><th>Percentage</th><th>Count</th></tr>
<tr><td>#1 || |PT:LOC| || #2 || |d| || #3 || || ||</td><td><i>20.08%</i></td><td><i>49</i></td></tr>
<tr><td>#1 || |PT:LOC/PRO3SG| || #2 || |d| || #3 || || ||</td><td><i>14.34%</i></td><td><i>35</i></td></tr>
<tr><td>#1 || |PT:PRO3SG| || #2 || |d| || #3 || || ||</td><td><i>9.02%</i></td><td><i>22</i></td></tr>
<tr><td>#1 || |PT:LOC/PRO3SG| || #2 || |l| || #3 || || ||</td><td><i>7.79%</i></td><td><i>19</i></td></tr>
<tr><td>#1 || |PT:DET| || #2 || |d| || #3 || || ||</td><td><i>7.79%</i></td><td><i>19</i></td></tr>
<tr><td>#1 || |PT:PRO3SG| || #2 || |l| || #3 || || ||</td><td><i>5.74%</i></td><td><i>14</i></td></tr>
<tr><td>#1 || |PT:DET| || #2 || |l| || #3 || || ||</td><td><i>3.69%</i></td><td><i>9</i></td></tr>
<tr><td>#1 || |PT:FBUOY| || #2 || |d| || #3 || || ||</td><td><i>3.28%</i></td><td><i>8</i></td></tr>
<tr><td>#1 || |PT:LOC| || #2 || |l| || #3 || || ||</td><td><i>2.46%</i></td><td><i>6</i></td></tr>
<tr><td>#1 || |PT:PRO3PL| || #2 || |d| || #3 || || ||</td><td><i>1.64%</i></td><td><i>4</i></td></tr>
</table>

Figure 4: Frequency view of PTs & orientation

Naturally, because of the systematic nature of IDglossing, sign types, be they non-lexical or partly- lexical signs, are able to be filtered through substring search matches to extract more specific hits. For example ^PT:PRO3|PT:PRO2 will find all third *or* second person pronouns (see Table 1).

	PT:LOC	PT:PRO3/PRO2
down	62%	58%
left	38%	38%
other	0%	4%
Total	100%	100%

Table 1: Results specifying for point type

There has been some discussion in the literature about the association of a downward palm orientation in pointing signs that are strongly associated with a location

[4] The six instances in which no sign precedes THINK are instances in which there has been a switch in hand dominance to the subordinate hand.

('here/there') and/or could be described as demonstratives ('this/that'), rather than being used simply pronominally. Even though the categorization of points in the Auslan corpus does not correspond neatly to the classes of pronouns, locatives, and demonstratives in traditional grammars, the data to date extracted from the Auslan corpus, of which the data in Table 1 is just an example, does not appear to show an association of a point with a palm turned downwards with at least locative meanings. It remains to be seen what a large reference corpus will show.

Lexical frequency by grammatical class

In the Auslan corpus there are annotations that assign grammatical class membership to sign tokens in context. In the multi-file multi-tier search dialogue lexical IDglosses can thus be constrained as co-occurring (overlapping) with grammatical class tags. The results can be view in frequency view and/or exported to databases for further sorting. Example results in Table 2 are based on two specific IDglosses, as shown:

	FINISH-FIVE %	FINISH-GOOD %
Adjective	5.10	0
Adverb	5.10	17.14
Auxiliary	36.74	31.43
Conjunction	2.04	5.71
Discourse marker	6.12	8.57
Interactive	1.02	0
Noun	3.06	2.86
Predicate	6.12	14.29
Unsure	1.02	0
Verb	**33.68**	**20.0**
Total	100	100

Table 2: The lexical frequency of two 'verbs' in the semantic area 'finish' specified by grammatical class.

The only major large lexical frequency study of any SL (McKee & Kennedy, 2006) did not, strictly speaking, take grammatical class formally into consideration in so far as it was assumed that the grammatical class of the English glosses used for each sign token accurately reflected each token's use *in situ*. In reality, glosses usually name the most frequent use of a sign, not its actual use in context.

Clause argument structure

In the Auslan corpus there are annotations that delimit clause boundaries. IDglosses are tagged for their status as arguments of the verb which is also tagged (e.g. as process, utterance or enactment). After merging tier annotations which combines these clause tags, it is relatively easily to identify and quantify clause construction types. For example, from the ELAN menu, > Tier > Merge Tiers, one selects tiers to merge (select 'concatenate'). View annotation statistics and select the newly created merged tier. Export to databases for further processing if necessary (a sample result from one file is shown in Table 3).

In the Auslan corpus there are also annotations that tag the identified overt arguments for their semantic role in the clause (e.g. as agent, patient, experiencer, etc.). By first merging the argument tag tier with the semantic role tier, before merging the result with the clause annotation tier, it is possible to extract richer data (Table 4).

Clause construction by order of overt arguments	**#**
V	27
A V	7
A1 V A2	6
V A	6
A1 A2	4
A	3

Table 3: Frequency of clause construction types

Clause construction by order of overt arguments	**#**
V (PROCESS)	27
A (AGENT) V (PROCESS)	6
A1 (AGENT) V (PROCESS) A2 (PATIENT)	4
A1 (CARRIER) A2 (ATTRIBUTE)	3
V (PROCESS) A (PATIENT)	3
A (ATTRIBUTE)	2
A (EXPERIENCER) V (PROCESS)	1
A (UTTERANCE)	1
A1 (AGENT) V (PROCESS) A2 (GOAL)	1
A1 (ENTITY) A2 (LOCATION)	1
A1 (EXPERIENCER) V (PROCESS) A2 (SOURCE)	1
V (PROCESS) A (ENTITY)	1
V (PROCESS) A (LOCATION)	1
V (PROCESS) A (UTTERANCE)	1

Table 4: Frequency of clause construction types specified for semantic role of argument

The data in Table 4 are only indicative of the type of information that can be extracted regarding clause structure based on secondary processing and are only taken from a single annotation file. Of 201 movie clips that have currently undergone primary and secondary processing, less than 10 have also been annotated for clause boundaries, overt arguments *and* semantic roles.

Though the range of clause construction types and the possible alignments of semantic roles to various argument positions commonly found in Auslan already appears much wider than that shown in the example file above, it is far too early to draw firm conclusions. A formal report of this data and its possible significance in describing grammatical structure in Auslan is not planned until the clause annotation set reaches at least 50 files and/or several thousand clauses.

Verb type by modification and by CA co-occurrence

In the Auslan corpus there are annotations that delimit periods of constructed action (CA). Multi-file searches constrained by values over three tiers can thus be based on the co-occurrence of tags for grammatical class (verb type), spatial modification (present or absent), and constructed action (present or absent). The results can then be exported to database programs. Relevant metadata regarding text type, age, and region, for example, can be easily appended to each token/hit in the exported data as ELAN automatically appends the file name source of each. This can then be run through statistical programs to test for factor interaction and significance.

Further value-adding: tertiary processing

The observations relating to single sign or multi-sign constructions extracted from a corpus using the procedures exemplified above are valuable in their own right. However, there is another, perhaps overlooked benefit to this type of SL corpus linguistics. The findings extracted from a corpus can themselves, in turn, be fed back into the corpus annotations, as part of an augmented secondary processing. They can then be used to generate yet further observations. I refer to this augmenting process here as *tertiary processing*.

For example, the very identification of a set of extremely high frequency lexical verbs in Auslan was only made possible because the corpus was not only annotated, but annotated in a systematic way that identified lemmas and, later, their grammatical class in context. The lexical frequency of sign types was then able to be added to IDglosses, filtered by grammatical class, as a frequency tag. In other words, researchers were able to find all instances of an IDgloss with a given grammatical class tag and replace that gloss with a tag signifying the lexical frequency of that sign (e.g., VHF for 'very high frequency', HF for 'high frequency', and LF for 'low frequency').

Augmentation of an exported data in this way can be done semi-automatically in database programs by filtering records and adding tags in fields for the relevant subset of records. The tag can then be added as another factor in subsequent re-evaluation of the data.

Inserting these tags into the ELAN file itself is worthwhile because there is currently a three-tier limit for simultaneous constraints in multi-tier multi-file searches. This means that any constraint which is itself the product of condition matching over two or three tiers, cannot itself be constrained further. By inserting such a derived value into the ELAN annotation file, this automatically means that this value can be used freeing the other query tiers to specify additional constraints.

Though the replacement or tagging process is not automatic within ELAN, there are workarounds. They take some time to do but since they need to be done only once and the results are always available for use, they are worth the effort (but see implications below). For example, the IDgloss tier can be copied or filtered to a new tier designed to hold the frequency tags. Then, the glosses on the derived tier are searched and replaced with the appropriate tag according to the lexical frequency by grammatical class table that has been generated by prior analysis. This can be done across multiple files, if not the entire corpus, in one operation. The workflow moves from the very high frequency signs to low frequency signs, as the very high or high frequency sign types are relatively few in number. (Of course, there are many tokens of these types!)

In other words, first with respect to high frequency signs, all IDglosses for a particular lemma are replaced with the same tag on the assumption they are all of the same grammatical class as the most frequent member. Then the remaining members of different grammatical classes—a much smaller set—are identified and the tag changed accordingly. With respect to low frequency signs, they can all be tagged as low frequency in one single universal search and replace: "find all annotations on the relevant tier which are *not* VHF or HF and replace with LF."

Similarly, as mentioned above, it is relatively easy to extract occurrences of signs that co-occur with periods of constructed action in a text. Tags for co-occurrence can then be added to the IDglosses (according to grammatical class).

Both frequency and CA co-occurrence information have been incorporated into a subset of the Auslan corpus in ways described above and were used in the recent study by de Beuzeville, Johnston and Schembri (2009) on the spatial modification of verbs in the Auslan corpus. This study examined the frequency and linguistic environments of verb modification with a view to assessing if spatial modification to signal these roles was obligatory in the language. The spatial modification of verb signs in SLs has traditionally been explained as a grammatical system marking subject and object roles (e.g., Sandler & Lillo-Martin, 2006), similar to obligatory subject marking in English (e.g., third person singular *–s* in *he walks*).

The Auslan study found that the modifications were not obligatory, were strongly associated with a very small number of high frequency verbs, and tended to co-occur in specific linguistic environments (e.g., co-occurrence with constructed action). The authors suggested that these observations would not be expected under the traditional grammatical account of spatial modification and are more in keeping with an analysis that sees the phenomenon reflecting, in part, the fusion of gestural pointing into the articulation of lexical verbs, as suggested by Liddell (2003).

It is anticipated that similar procedures as those described here will integrate derived clause argument structure patterns into tags added to clause annotations within the Auslan corpus. The patterning of clause chains (e.g. with overt or elided arguments, or with certain verb/argument sequences) and their interaction with verb modification, depicting signs, constructed action (as well as other linguistic variables) may then become identifiable and amenable to quantification and further analysis.

Standardizing annotation schemas

The type of investigations of the Auslan corpus that we have briefly illustrated here have only been made possible because of the distinctions made in the IDglossing between types and tokens, and between sub-types of signs. The way these distinctions are coded in the in the IDglosses are described in detail elsewhere (e.g. in corpus annotation guidelines[5]). The primary, secondary and even tertiary processing of language corpora is extremely time consuming work. However, the results more than justify the effort expended in adding value to raw language recordings — recordings which would otherwise be of limited use — in this way.

The international standardization of annotation practice, protocols or schemas is highly desirable. Indeed, at the level of primary processing this should be a high priority. At the level of secondary processing, however, there is much more room for flexibility as the aims of various research teams can be very different, each perhaps requiring its own dedicated secondary tags. Standardization, in so far as it is possible, will certainly enable the corpus-based comparative analysis of SLs to be undertaken.

Within a give SL corpus, however, there is really no option: standardization in terms of systematicity and consistency is mandatory. Only in this way can annotations create machine-readable SL texts that can be searched rapidly and with great precision. The results can then be further processed for statistical significance and interaction, or, just as importantly, the hits further examined individually *in the media context* to assist in the determination of their semiotic or linguistic significance.

Implications for annotation software

From the discussion above, it will be evident that the steps needed to conduct some searches or data exports are in need of automatization. For instance, preparations for some multi-tier pattern match searches, on the one hand, or merging information coded on separate tiers, on the other, are ad hoc and time consuming. External plug-in scripts are one solution. However, fully integrated improved program functionality is preferable as it means all researchers using the same software have the same functionality available.

With respect to ELAN, for example, these scripts or routines would enable one to automatically create, copy or merge certain tiers in multiple annotation files of the same type; automatically look up an alternative value for an annotation in a table and substitute that value for the annotation on a particular tier in multiple annotation files; or automatically place a specified value in an empty annotation field which is the result of a hit specifying the overlap of two annotations of two other tiers (independent or otherwise).

Search functionality also needs to be improved so that more than three tiers may be specified in constrained pattern matching. Most importantly, the co-occurrence (or non-occurrence) of two given annotations within the time delimitation of a single annotation on another tier should able to be specified as a search condition.

Conclusion

The creation of SL corpora as corpora in the modern sense involves more than recording, digitizing, editing, cataloguing and archiving video texts. Corpus creation must also involve the transformation of archived material into something which is machine-readable by the principled application of annotation procedures that make optimal use of new digital technologies. By adding value to a corpus through systematic and principled primary and secondary processing, it is possible to extract the true value inherent in a linguistic corpus.

Acknowledgements

Research towards this paper was funded by an Australian Research Council Discovery Project grant (#DP1094572) awarded to Trevor Johnston.

References

Beal, J. C., Corrigan, K. P. & Moisl, H. L. 2007. “Taming digital voices and texts: Models and methods for handling unconventional synchronic corpora”. In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds), *Creating and Digitizing Language Corpora Volume 1: Synchronic Databases*. New York: Palgrave Macmillian, 1-16.

De Beuzeville, L., Johnston, T. & Schembri, A. 2009. “The use of space with lexical verbs in Auslan: A corpus-based investigation”. *Sign Language & Linguistics*, 12 (1), 53-82.

Johnston, T. 2010. From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics, 15*(1), 104-129..

Johnston, T., & Schembri, A. 2010. Variation, lexicalization and grammaticalization in signed languages. *Langage et Société, 131*(mars 2010), 19-35.

Liddell, S. K. 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge: Cambridge University Press.

McKee, D., & Kennedy, G. (2006). The distribution of signs in New Zealand Sign Language. *Sign Language Studies, 6*(4), 372-390.

MPI/LAT Technical Group: (Head) Wittenburg, P., (Team members) Auer, E., Broeder, D., Gardellini, M., Kemps-Snijders, M. et al. 2009. *EUDICO Linguistic Annotator (ELAN)* (Version 3.8). Nijmegen, Netherlands: Max Plank Institute for Psycholinguistics: Technical Group (Language Archiving Technology). Available at: http://www.lat-mpi.eu/tools/elan/.

Sandler, W., & Lillo-Martin, D. (2006). *Sign language and linguistic universals*. Cambridge: Cambridge University Press.

[5] The Auslan annotation guidelines can be downloaded from http://www.auslan.org.au/about/annotations/