On Syntactic Structure Representation

(author's summary – v.1.7.6)

A.Hayrapetyan. On Syntactic Structure Representation (in Eastern Armenian: *Բնական խոսքի ընդհանրական ներկայացման մի տարբերակի մասին*). Agoulis, Concord, 2022

Table of Contents

1. Foreword

1.1 Annotation

Humans create images of reality in their mind using the input from senses. They describe, externalize these images, in particular, using speech. Production and comprehension of speech is a conscious, cognitive process. The generation (production) and analysis (consumption) of speech occurs according to the set of rules of natural languages. Language and speech are related as producer and product. The syntactic structure of speech is universal for all natural languages - it is a tree like structure of signs denoting processes, concepts and objects of reality. An approach to designing such a structure – the context tree - that can be produced and consumed using the rules of any natural language is discussed in the book.

1.2 Structure of the book and the summary

The book contains 322 pages, 5 drawings, 136 tables, 198 glossary terms. There are 90 references cited in the work. The index contains about 1200 terms.

The 2 subsections below repeat the content of the original namesake sections almost entirely: only few paragraphs at the end are skipped. The sections following the “Foreword”

cover: a) NLP (Natural Language Processing) basics, b) Eastern Armenian morphology (*Word forms and morphemes*, *Nouns: declension*, *Verbs: conjugation*, and *Summary of morphology* sections of the original), c) syntax (*Sentence*, *Intellectual Activity*, and *Formation of Meaning* sections of the work) d) content of the Appendices, e) Conclusions, and f) Next Steps. The last two are not part of the work.

1.3 Purpose of the work

This is a multi-purpose work:

1. The content of the speech (oral or written) is represented as a Content Tree (CT), which is a tree of signs (words) denoting concepts or proper names. The CT has various applications: multilingual search engines, translation systems, encyclopedic and linguistic (monolingual, bilingual, multilingual) dictionaries.
2. Another goal is to outline the principles of computer model implementation for Armenian language and to list some applications of the model.
3. It is also a database of Eastern Armenian morphological and syntactic information, which can be used for NLP and alternative implementation of Armenian language model.

The CT is built algorithmically according to the morphological and syntactic structures of given language using the dictionaries of morphemes and wordforms. A system of such algorithms and dictionaries for different languages allows transforming the text written in one language into a set of CTs, interpreting them using algorithms and dictionaries of another language, thus reproducing the original text in different language. However, the CT alone is not enough. The goal of translation is communication of meaning rather than content. During translation, the meaning in one language coded by the CT is represented by the Sentence Tree (ST) of the other language, which becomes a sentence after being linearized.

Examples of applications include automated corpora creation, comparative linguistic studies, preparation of texts for publishing, etc.

We describe the approach using the Eastern Armenian language. The morphological model specified in [Jah1974] is used as a base for stemming algorithm – the foundation for building computer model for Armenian text analysis and production. The approach is applicable to other, in particular, Indo-European languages.

We map the elements of the Armenian language into the elements of universal model and vice versa. The set of categories and elements of a particular natural languages are accumulated into the categories and elements of universal language model.

The language model is described as formally as possible. However, formalism is not a major goal: the intent was to make definitions of concepts, terms, and their relationships equally convenient for humans and machines (algorithms) alike. The model is designed for text analysis and for text generation.

Language is a biological phenomenon and, hence, it is hard to express it by formulas or equations: it is difficult to formalize it. The exceptions to rules and to regularities that are missing in the book are either minor or the author is unaware of them.

1.4 Structure of work

The first part (chapters 2-3) is dedicated to the role of thinking and language and their relationship. The role and basic features of language and speech, in particular, the meaning, the

content and the expression plains of speech, the general structure of the sentence, natural language structures are examined.

The second part (chapters 4-7) is devoted to the morphology and typology.

In the third part (chapters 8-11), based on the analysis of regularities and exceptions in morphological data, the text forms generation (using tree-like representation of the paradigms) and text forms analysis: stemming, tagging, and lemmatization algorithms are designed. The design and description of the morphological model is summarized.

The fourth part (chapters 12-13) discusses phrase and sentence structure and summarizes the approach to creating a computer model for the universal representation of natural speech. The notion of verb signature is an extension of valency. By using verb signatures the declension system of Eastern Armenian is analyzed, the alternative declension table as well as a declension of synthetic wordforms are described. An algorithm based on the analysis of syntactic rules of Armenian and other languages, for building the CT is outlined.

The main goal of the suggested universal representation of natural speech – the CT - is to develop a universal structure that can a) represent a speech expressed in any natural language and b) reproduce the speech represented by CT in any other language. The content-meaning mapping is examined (chapter 14).

Apart from being a systematic representation of speech and having various applications (chapters 15-16), CT is a tool for analyzing natural language and speech.

The appendices are intended as reference data. The information about the word forming inflections, the systems of declension and conjugation, the examples and other morphological and syntactic facts are presented in tabular form.

2 Introduction

In introduction the relation of thought to speech, as well as the sign (signifier) to the signified is discussed.

Humans create internal images of the surrounding reality by thought, which they express by speech. Speech is constructed according to the laws of language. Language is a set of means for producing speech: the terms (morphemes, words, word groups) and grammatical rules that use these terms to express the content.

A word acquires meaning through context (in a broad sense). The formation of meaning - making sense, is beyond linguistics. Logic is also beyond linguistics. However, since ancient times, grammarians have dealt with both: the logic and the semantics, often merging them with grammar.

The Stoics [Bob2006, Chapter 5.] and G. Frege [Fre1892] distinguished 2 meanings of the word *meaning*. "The late nineteenth-century mathematician-philosopher Gottlob Frege provided a concise distinction between these two often-confusing aspects of the word meaning. He distinguished between the *sense* of the term and its *reference*. Its sense is the idea that one has in mind that corresponds with considering a particular word or phrase. This is distinguished from the reference of the same word or phrase, which is something in the world which corresponds with this term and its sense" [Dea1998 ::61]. We will refer to this distinction here

only to emphasize that meaning is outside of the linguistic realm and that grammar is about content, rather than meaning.

In a sense the content is the intensive description of something while reference is the extensive description of the same thing. Meaning is the mapping (matching) of these two descriptions.

Thus, a sentence should at least be about a state (or an object) and reaching or staying in the state. If either of these is absent then the sentence is incomplete.

The content of the complete sentence is understandable. It tells *what* (*who*) is [doing] *what*.

In the middle of the previous century, Chomsky developed the hypothesis of autonomy of grammar, by which he removed the meaningfulness of speech from the requirements to grammar.

The "Colorless green ideas sleep furiously" [Cho1975::138] is an example of meaningless, but grammatically absolutely correct sentence. Explaining the history of this often quoted example in conversation with Mitsou Ronat he says: "I tried to show that every clear formulation of a hypothesis concerning the alleged necessity to define syntactic notions in semantic terms led to incorrect results. Thinking about these question led to what was leter termed the hypothesis of *autonomy of syntax*" [Cho1975::138].

Assigning meaning to sign is a much more complex process [Fre1892], [Pei1998::98], [Hof1979::82], [Dea1998::59-101] than just simplistic, static relation of the signifier (the sign) to the signified (the object).

3 Natural language model

3.1 Phrases

The meaning of the speech is encoded through the content in units of different levels: tone (stress), phoneme groups (syllables), words, word groups, and phrases.

Words and word groups, as the minimum components that make up phrases, are classified into three main categories: objects or states (noun), change of state (verb) and attributes (adjective and adverb).

"The statement that part of speech is a group of words should be taken tentatively. If we abandon the view that the real word is a grammatically formed wordform (whether it has a synthetic or analytic structure) and that the classification of parts of speech should be a classification of minimal syntactic units (E. Atayan), then many syntactic units that are considered as part of speech are not words, but rather groups of words (e.g. *to have made to write* (verb), *as if* (adverb), *not only .. but also* (conjunction), etc.). Thus, the concept of lexical units should be distinguished from the concept of words. Parts of speech are not groups of words, but rather groups of word-like units" [Jah1974::133].

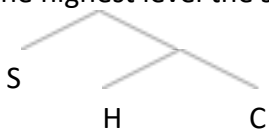
The sentence is the most complex unit for expressing content. It is composed of these phrases (P):

1. Noun phrase (NP) - the main member in this structure is a noun or a pronoun or any part of speech used as a noun.
2. Verb phrase (VP) - the main, independent member is a verb.
3. Attributive (Adjective) phrase (AP) - the main member can be a direct or oblique form of the noun or adjective, as well as a participle or adverb.

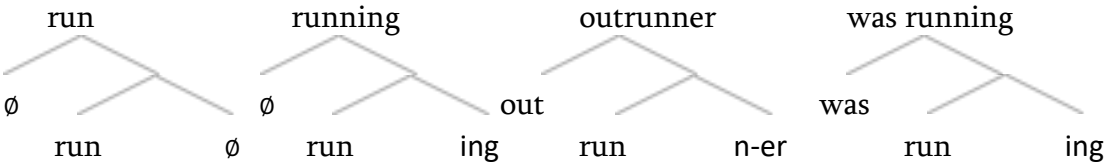
3.2 Speech structure universality

When we look at human language as a universal phenomenon (as opposed to particular language), we notice that the speech covers a wide spectrum of forms, starting from completely formal mathematics and programming languages, through scientific, administrative, or everyday styles of speech up to the art, where we find abstract and absurd literary forms.

At the highest level the sentence has this universal structure – the linguistic tree:



The letters of the nodes mean: S - specifier, H - head, C - complement. Phrases [Dev2000::297-307], morphemes, words, word groups can also be described by the linguistic tree. For example, these trees represent lexemes built by *run* as the head:



3.3 Types of grammatical laws

There are 2 types of rules for sentence or text wordforms composition, i.e. grammatical rules:

1. Mandatory (requiring) - enforcing or prohibiting.
2. Optional (allowing) - not enforcing and not prohibitive.

3.4 Model

3.4.1 Terms and concepts of the model

The model consists of these algorithms:

1. Wordform generation - generates dictionary forms (lemmas) from simple (root) stems and wordform suffixes and their derived forms using paradigmatic suffixes
2. Stemming – extracts simple stem and suffixes from the text wordform
3. Lemmatization - calculates the dictionary form from the text wordform
4. Tagging - determines the part-of-speech and paradigmatic category of the text wordform and tags it accordingly
5. Syntactic analysis - analyzes the phrases and other parts of a sentence and determines their hierarchic relationships
6. Transformation – creates ST based on the data of previous step and transforms it into CT and back
7. Sentence generation – creates Armenian sentence from the ST.

The model is described using traditional linguistic concepts: a) morphological (in a broad sense): sentence, phrase, lexeme (word or group of words), [simple] base, suffix, particle, b) grammatical: class, case, number, determinacy, voice, aspect, mood, tense, person, c) parts of speech: noun, adjective, numeral, pronoun, participle, verb, adverb, adposition, conjunction, modal and natural sound words, d) syntactic: subject, verb, predicate, object (complement) and

words and structures that are not part of sentence (greetings, vocative forms, pro-sentences, etc.).

"The fact that despite much of criticisms and rebuking, rejection and replacement attempts, classifications and reclassifications, the concept of parts of speech remains one of the central concepts of modern grammar, speaks for its being an objective reality" [Jah1974::125]. This thought is probably true for other linguistic concepts and notions, because they are formed by and survived under the centuries long pressure through cultural evolution.

The roles of lexemes in the sentence: attribute, predicate, arguments (subject, objects), which are expanding into attributive, verb, and noun phrases, are determined by part of speech.

3.4.2 Model Quality

The generative laws can be described by a variety of mathematical means: formulas, tree-like structures, networks, production rules, etc. Regardless of the representation, the rules in the model can produce:

1. All correct forms, but do not guarantee that every derived form is correct;
2. Only correct forms, but do not guarantee that all correct forms are generated;
3. Only all correct forms, i.e. meet both conditions: a) any generate form is correct (unlike #1), and b) there is no correct form that the model does not generate (unlike #2).

This triplet of quality specification is applicable to other algorithms too: stemming, tagging, searching, etc.

3.4.3 Model Coverage

A complete language model typically covers the following areas of linguistic study:

1. Study of expression – patterns of vocal (phonemes), written (graphemes), and gestural (gestemes) expression of speech.
2. Morphology – study of patterns and structures for generation, analysis, and transformation of words and word groups.
3. Syntax - determination of content expression components and their correct combinations.
4. Semantics - expression of meaning by signs and symbols (wordforms).
5. Pragmatics - analysis of the different ways and styles of expressing the same meaning.

The developed model is limited to the second and third bullet-items.

One of the goals of developing the model is to define a generic description of content that can be expressed (or rather interpreted) in any language. Such description of content allows to:

1. Map and store the content expressed in any language in a generic form.
2. Express the stored content in any language.

Such model allows translation from and to any language if the above 2 functions are implemented for these languages.

A sentence generating system actually transforms one sentence into another sentence.

4 Morphology

4.1 Categorization

These are the types of Armenian lexemes: 1) noun, 2) adjective, 3) numeral, 4) pronoun, 5) verb, 6) adverb, 7) connection, 8) conjunction, 9) modal words, 10) natural sounds (including

exclamations): Numerals do not have any special grammatical features: they behave either as nouns or adjectives.

The first four types undergo declension, while the verb - conjugation. Some participles in addition undergo declension.

In the *Noun Forms: Declension* and the *Verb Forms: Conjugation* sections the components and the structure of Armenian parts of speech are analyzed. The 18 [Jah1974::189-196] patterns of external: I, U, An, Voj, Va and C and internal: O and A declension paradigms as well as regular A, E, and irregular conjugations are described. The stem formation for 5 common and 8 imperative verb specific patterns of are examined.

Determiners (articles) and plurality are considered part of the declension system.

These cases (slightly different that currently accepted) are identified by syntactic data analysis (the initial in square brackets is the case tag):

	case	Armenian	Russian	alternative
1	[N]ominative	*Անվանական*	Именительный	Absolutive
2	[G]enitive	*Սեռական*	Родительный	Possessive
3	[D]ative	*Հանգական*	Дательный	Allative
4	[O]bjective	*Իրական*	Объектый	Direct
5	[A]blative	*Բացառական*	Исключительный	Delative
6	[C]omitative	*Ուղեկցական*	Сопроводительный	Instrumental
7	[L]ocative	*Ներգոյական*	Местный	Propositional

The verb forms derived from the same root can differ in kind (transitivity), voice, tense, mood, form, person and number. Transitivity and voice generally do not have morphological markers for determining the attribute - they are explicitly defined in dictionaries and do not change during transformations in four-dimensional: tense, mood, person and number, space of conjugation. Aspect changes can be hardly considered independent or separate. They are inseparable from tense and mood.

These are Armenian Participles:

ID	participle	tag	Armenian	example	Arm. example
1	infinitive	INF	*Անորոշ*	[not] to run	[չ]*վազ-ել*
2	present	CONT	*Անկատար*	[not] to be running	[չ]*եմ վազ-ում*
3	synchronal	SYN	*Ժուգընթացական*	while [not] running	[չ]*վազ-ել-իս*
4	prospective	PRSP	*Կատարելի*	will [not] run	[չ]*եմ վազ-ել-ու*
4a	predicative	PRED	*Ստորոգելի*	runnable	*կատար-ել-ի*
4b	future	FUTP	*Ապառնի*	to be ran	[չ]*վազ-ել-ի-ք*
5	past	PSTP	*Վաղակատար*	have [not] ran	[չ]*եմ վազ-ել*
6	resultive	RSLT	*Նախընթացական*	ran	*վազ-ած*
7	progressive	PROG	*Համընթացական*	running	*վազ-ող*
8	negative	NEG	*Ժխտական*	not run	*չեմ վազ-ի*

Participles that are not used independently (with no auxiliary verb) are preceded by an auxiliary verb (first person, singular) in the examples.

The Simple verb forms in Armenian are:

ID	form	tag	Armenian	example	Arm. examp.
9.{1-6}	preterite	PTE	*Անցյալ կատարյալ*	have [not] run	[չ]*վազ-եց-իր*
10.{2,5}	imperative	IMP	*Հրամայական*	run	*վազ-իր*
11.{1-6}	subjunctive future	SBJ.FUT	*Ղեձական ապառնի*	would [not] run	[չ]*վազ-ես*
12.{1-6}	subjunctive past	SBJ.PST	*Ղեձական անցյալ*	would have [not] run	[չ]*վազ-եիր*

In Armenian examples the second person singular form of the verb is given. The verb forms numbers in parentheses mean: 1-3 are the singular 3 persons, and 4-6 are the plural.

The moods of Armenian are: 1) indicative, 2) imperative (including prohibitive), 3) subjunctive, 4) assumptive, and 5) necessitative.

The peculiarity of conjugation is that most of the lexeme forms are used in word groups: a combination of the auxiliary verb, particle, and verb form.

The connectors and conjunctions can also be represented by word groups.

The structure of the declension and conjugation forms are defined by BNF-like generative expressions.

From single noun dictionary form 5 singular and 5 plural, 10 possessive per first and second person and 2 determinative forms are derived, totaling 34 text forms overall. This number can be less, for example, in the absence of a locative case, plural or article form, but sometimes it can be more.

From the verb dictionary form – the infinitive – these text forms are generated:

1. 8 participles, 12 past perfect (6 positive and 6 negative), 4 imperative and prohibitive, 24 subjunctive (6 for each: present, negative present, past, negative past) and 12 necessitative forms:
2. The above-mentioned 60 forms are specific to the vast majority of verbs, but there are verbs that do not have some forms (for example, present – it is the same as synchronal) or have two or more variants of others (for example, the imperative):
3. Infinitive, accusative, past and present verbs are also declined: the first by U and the others by I declension.

4.2 Word structure

A classification of morphemes is proposed, according to which 19 slots are needed to combine text forms from 4 types of prefixes, 1 simple stem, and 14 types of suffixes.

4.3 Composition of the dictionary storage

The main part of the model is the dictionary of terms: simple stems and suffixes. They are described by a number of attributes, including a tag, which defines the type of the form.

The dictionary is subdivided into sections according to the 19-slot structure of words: a) prefixes (4 subsections according to positions), b) suffixes (14 subsections), c) immutable and simple words (further divided into subsections: common, verb, and natural sound stems), d) proper names (and bases), e) lexemes (word groups).

These are Armenian simple stem types

	tag	description	example
1	STM.DUAL	Does not get aspect suffix	անվան (name)
2	STM.IMPF	Imperfective stem, receives perfective suffix	խաղ (play), հոգ (care)
3	STM.INDT	Receives either Inperfective or perfective suffix	մեծ (big), ձանձր (bore)
4	STM.NAT	Natural sound stem	դխկ (dkhk), փսփս (psps)
5	[STM.]NOML	Receives nominal suffixes	անվան (name), մտ (mental)
6	[STM.]NOUN	Receives case suffixes	խաղ (play), հոգ (care)
7	[STM.]NUM	Receives numeral or adjective suffixes	չոր (four), յոթ (seven)
8	STM.PERF	Perfective stem, receives imperfective suffix	մտ (enter), տար (carry)
9	[STM.]PRON	Receives case suffixes	ոչնչ (illuminate), այնտեղ (there)
10	ADJ	Receives nominal and case suffixes	այլախոհ (dissident)
11	ADJR	Receives superlative suffixes	մեծ (big)
12	ADV	Receives nominal and verb suffixes	արագ (fast), դանդաղ (slow)
13	IMP	Imperative	տար (carry), տես (look)

Some affixed and compound stems are added to the list of simple stems, which includes unalternated as well as alternated stems. The affixed forms have irregular inflections. These are irregular plurals, non-singular and non-plural forms, which can be dictionary or text forms.

4.3.1 Tag types

Tags are divided into these types and subtypes:

1. Part-of-speech - parts of speech and their narrower groups, for example ADJ, NAM
2. Paradigmatic – case, number, tense, aspect, for example: IMP, PLU, L.
3. Grammatical - indicating grammatical categories, for example: IMPF, RFL.
4. Morphological - characterizing morpheme types, for example: STM (stem), POX (suffix), INST (preposition).
5. Operational - T (terminal), NT (non-terminal), etc.

The last three usually do not appear in the textual form marking; they are used for calculating tags.

There are two matrices, which define compatible and incompatible pairs of tags (morphemes).

4.3.2 Description of morphemes

The morpheme description in the dictionary (catalog is probably more accurate) consists of tags and constraints specific to the form (general constraints are not included). in the stems definition forms there are ontological relations, meaning (sense), paradigmatic trees, verb signatures, etc.

There is a field, which indicates the morpheme that can be a part of a word group. This field helps with identifying the word group in text and finding relevant entry in the lexemes section of the dictionary.

4.4 Generation of text word forms

For a given stem or dictionary form the system generates:

1. All text forms:
2. All paradigmatic forms (for example, all forms of the noun, all personal and impersonal verb forms of the infinitive).

3. The form specified by grammatical features (for example, the Dative plural with the possessive article of second person).

Particular variety of a paradigm is defined by tree-like structures. These structures are listed in the dictionary. The trees corresponding to a morpheme or a lemma are specified in the dictionary description.

4.5 Word form analysis

The purpose of the three analytic algorithms: stemming, tagging and lemmatization, is to determine the part-of-speech, inflectional type, and the lemma of the textual form.

5 Syntax

The content of sentence contains information not only about the relationship of things, but also the speaker's attitude to things and their relations, as well as evaluation of the sources of information communicated. In general a sentence communicates information about:

1. the things, positions and states and their relations.
2. the changes in relations of things, positions, and states.
3. the relationship of things, positions and states and the speaker's attitude towards their changes.
4. the sources of information about the relations of things, positions and states, how the knowledge is obtained and how surprising the information is.

5.1 Verb (sentence) signatures

The [ordered] sequence of word forms in a phrase and the position of the main form is called phrase signature. For example, the series of arguments of a verb and the position of the verb in that series is the signature of the verb [phrase]. A verb can have several signatures. Since the verb phrase is basically the sentence (see the [Jah1974::332 citation below], its signature is actually the signature of the sentence. For example, the signature of the verb *to imply* is N+...+D (dots indicate the position of the verb). It means that the verb *to imply* aligns with the subject in Nominal case before it and requires an object in Dative case.

"If the expansion of nominal categories leads to a nominal phrase, then the expansion of verb or predicate categories ultimately results in a sentence" [Jah1974::332]. The set of verb signatures defines all possible sentence structures of a given language. The verb is the axis around which the verb arguments: subject, predicative, objects with their attributes, as well as adverbs, are grouped.

The category of valency, which is pertinent to verb, determines the number of its arguments. The verb signature, in addition, defines their inflectional type and relative positions.

If valency is an extension of the notion of transitivity, then the signature is an extension of the notion of valency. If the valency defines the number of verb arguments, then the signature defines the types of arguments (vowels) and their relative positions.

Verb signatures are used by sentence parsing algorithms and are included in the list of characteristics describing verb forms in the dictionary.

5.2 Unified Declension

In natural languages, the relations of things (objects) are expressed by two classes of cases: morphological and semantic (for alternative names of classes, see: Has2006). The former

indicate the subject or direct object such as Agentive, Active, Accusative, Objective (it is also in the semantic group), etc. cases.

Semantic set is the largest. These cases reflect spatial, temporal, ontological, etc. relationships.

The objects (spatial positions and time points are also objects) can be involved in activities at the beginning, at the end, or during the action.

If we consider case forms as positional (we are talking about the ontological relation to the action) signs of the objects that are involved in the action, then we can identify these common groups of cases:

1. Objective - action participants: agent or experiencer.
2. Allative (Dative) - object (including place and time) where the action ends.
3. Elative (Ablative) - object (including place and time) from which the action begins.
4. Comitative (Instrumental) - an object (including place and time) that accompanies (in a broad sense, simultaneous or parallel to) the action.
5. Locative (Prepositional) - an object (place, position, time) in relation to which (in or around) the action takes place.

Nominative and genitive cases are not included in these groups, because they define attributes of the action (event) or participants of the action: the nominative is mainly the determiner of the action, and the genitive is the determiner of the participants. The forms of these cases can be combined with adpositions and form composite forms that can be classified in one the above groups. The Genitive form has all the properties of a relative adjective, but it can also be combined with adpositions and form another case.

5.3 Unified Conjugation

The unified conjugation system includes all theoretically possible tenses, aspects, voices, as well as modes available in natural languages. Theoretically verb categories of any language of can be uniquely mapped into the categories of unified conjugation and vice versa. However, additional research (in particular, on the interdependence of tense, aspect, and mode) is necessary to evaluate the convenience and the benefits of such mapping.

Even if we accept that it is no reasonable mapping, we can still include the verb categories from all languages in unified conjugation system.

5.4 Analysis

5.4.1 Components of Speech

Sentences and extra-sentential components that comprise speech can be identified by punctuation.

"Thus, examining the sentence means examining the following five categories components: 1) modal words and word groups that denote mood, 2) words and word groups that denote relations, 3) words and word groups that denote subject, 4-5) words and word groups that denote objects: the latter depending on levels of connectedness to verb can occupy different positions in relation to it. They are traditionally known as predicative, object and adverbial. They are the so-called secondary members of the sentence (we put the predicative here with some reservations as the "object of the substantive verb")" [Jah1974::337].

A sentence is a subordinate-recursive structure of phrases, which at the lowest level is a chain of lexemes.

We distinguish: a) speech components - sentences (simple sentences) and modal, relational, and vocative structures, b) sentence members: subject, predicate (verb), etc., and c) lexemes: noun, verb, etc. from which the speech components are built.

5.4.2 Constructing a sentence tree

"Speaking a language involves transforming structural order into linear order and conversely, understanding a language involves transforming linear order to structural order" [Tes2015::12] (according to Jia2015).

The sentence parsing algorithm using the results of stemming and tagging and the relative position of the lexemes in the sentence, determines:

1. members of the sentence and their roles (in some cases clarifying the ambiguities of stemming and tagging):
2. interdependencies of sentence members (to build the sentence tree).

After passing through the stemming, tagging, and lemmatization algorithms, each word of the sentence gets tagged and its lemma is restored. Based on the tags and the relative positions of the words in the sentence, the ST is built.

5.4.3 Building the content tree

To build a CT from a ST, the unified paradigm labels are assigned to the nominal and verb forms, and the lemmas are replaced by general identifiers (GIDs), which denote the meanings of lexemes.

Such a CT should be "understandable" to the algorithms that generate sentence in other languages.

5.5 Formation of meaning

5.5.1 The complexity of meaning formation

The meaning is encoded by the structure of morphemes, lexemes, phrases, sentences, and the entire speech event. However, apart from the form, it also depends on the context of each of the forms. In other words, the meaning of any form depends on both: its structural components and the whole structure which it is a part of.

The principle of autonomy of syntax does not mean that there is no connection between form and meaning. It means that the laws of grammar are aimed at expressing content. The meaning is attached to a particular form dynamically depending on the context and the speech event.

Grammatical laws often depend on the meaning of the terms. In order to make the correct paradigmatic transformations, you need to know whether the form is a noun (at times the class too), a verb, an adjective, an adverb, etc. in what sense it is used. For example, the word *avel* (*more* or *broom*) is a verb by form - an infinitive (infinitives in Armenian end either with -el or -al). Only after determining that the stem av- is meaningless, it becomes clear that the laws of conjugation are not applicable to that form. The meaning of the word *avel* depending on the context can mean *broom* or *more* – we can decide whether grammatical rules relevant to noun or adverb can be applied only after analyzing the context.

For encoding the meaning at each subsequent level of the syntactic structure, that is, for building correct structure, for applying relevant grammatical rules, we need to determine the meaning of the structure at the previous levels. This is the reason why structural or semantic

approaches [Jah 1969::85-91] to linguistic modeling separately are not sufficient for speech analysis.

5.5.2 Sentence transformation

The purpose of speech is to convey meaning rather than content. In order to determine sentence-meaning correspondence, it is necessary to find out 1) which sentences express given meaning, or 2) what meaning is expressed by a sentence. The first is the task of translation: matching the same meaning to sentences in different languages. In other words, translation is a modification of sentence structure that preserves meaning. Meaning is the invariant of translational transformation.

In general, the sets of signs (lexemes) and concepts are different in different cultures. The ST expressing the same meaning differs not only by the signs assigned to nodes of the tree, but also by its structure. The same or different meanings can be expressed by the same or different word groups or phrases. The sign – concept mapping is culture dependent.

When translating speech from one natural language to another, it is not enough to separate lexemes in the input speech and replace them with lexemes of the another language. It is necessary to consider sentence, or rather, the speech context in whole.

6 Implementation

In the most general case of synthetic language the system include:

1. Databases:
    a. Dictionary of Simple stems
    b. Dictionary of immutable forms (and particles)
    c. Dictionary of morphological suffices
    d. Dictionary of paradigmatic suffixes
    e. Table of tag compatibility and constraints
    f. Thesaurus
    g. List of paradigmatic structures (morphological signatures of lexemes):
    h. List of verb [phrase or syntactic] signatures:
    i. Dictionary of proper names (and stems).
2. Algorithms:
    a. Forms Generator - builds text word forms (from a given lemma or a stem).
    b. Stemmer - determines the components of the word form.
    c. Tagger - determines part of speech and paradigmatic form.
    d. Lemmatizer – restores the direct (dictionary) form of the text form.
    e. Parser - builds ST.
    f. CT Transformer – converts ST to CT (maps particular language text forms to unified paradigm forms).
    g. ST Transformer – converts CT to ST (maps unified paradigm forms into text forms of a given language).
    h. Sentence generator - expresses content in a particular language: ST linearizer.

The proposed CT format ensures generality not by abstracting or ignoring the specifics, but on the contrary by taking them into account. It is not abstract-universal, but rather specific-cumulative.

7 Applications

The language model is the basis for a) creating thesaurus and b) encyclopedias, c) constructing corpora (treebanks), d) checking spelling and grammar, e) publishing, f) translation, g) search engines, and h) linguist's workbench.

8 Summary

People analyze the images received from the senses into ontological, relational and logical concepts. The environment (in a broad sense) is represented as relations of taken apart or chunked [Hof1979::382] together invented by humans concepts. These make up our perception of the universe and what we have identified with reality in our brains. This is the picture of reality.

Modern science still does not know how it is formed and stored in the brain. People judge about the image of reality by presenting, communicating to each other. Communication is done via speech, be it in natural (including mathematical and logical), pictorial (drawing, picture), or another type of language.

Natural speech, together with the senses, conveys information about the environment. It participates in constructing semantic network (SN) [Hof1979::370-372] in the human brain.

The construction of SN is done via 1) perception and 2) testimony:

1. [Sensory] Percept ⇒ Image ⇒ ((Consciousness)) ⇒ [[Thinking]] ⇒ Concept ⇒ SN
2. SN_s.Concept ⇒ ((Language)) ⇒ [[Speech]] ⇒ ((Language)) ⇒ Concept ⇒ SN.

In the above 2 cognitive processes the actions are in double square brackets and the sets of laws or constructs (as opposed to object) in double curved brackets (parenthesis). The speaker's SN is marked as SN_s to distinguish it from the SN of the listener (perceiver). The first, the perception process is a schematic description of thinking, and the second, the evidential (testimonial) is a process of communication. With the latter, people report to each other what they saw, heard, or think about the reality and events.

Speech is a sequence of linearized tree-like structures of symbols (signs), which encode concepts according to grammatical rules. The symbols can be simple - morphemes or complex - phrases. Concepts are denoted by linguistic trees of symbols. A concept sign is a combination of linguistic trees. This sign is complicated not so much because of the parallel branching, but because complexity of determining the meaning of the tree: one must go back and forth from sign to meaning at different levels of the tree, each time clarifying the meaning of partial tree (branch).

The set of concepts is mostly universal while the linguistic tree is exclusively universal, biological. However, the coding (denoting) of signs and concepts by linguistic trees is specific, cultural. In other words, the concepts and the principles of construction of their verb signs are the same for all languages, but the structure (realization) of a specific concept signs is different. The language organ formed in the brain is a set of rules, according to which we identify the structures of specific signs corresponding to the generic pronouns {*what*, *who*}, {does, is} *what* [by *what*] [*when*, *where*, *how*] in the universal communication structure - in the sentence - are identified.

The proposed CT should not be considered as a deep universal structure, but rather as a format for linguistic information exchange - a data contract.

The purpose of this work is to bring forward the importance of the API (of which the data contract is a part) implementation, to show the possibility of such implementation and to identify solvable and unsolvable (hard) problems.

9 Appendices

9.1 Information Appendix A

The types of morphemes and their marking (tags) of parts of speech, their subclasses, as well as common and verb categories, are listed.

Word formation suffixes are described along with the examples of their use.

9.2 Information Appendix B- Declension System

The classes and subclasses of Armenian nouns and the declension forms along with their meaning and roles in sentence presented in tabular form. Different paradigms of the Armenian external: I, U, An, Voj, Va, Ts, and internal: Vo, A declensions of are given. The case and plural suffixes are separately summarized in tables.

9.3 Information Appendix C- Conjugation System

The structural parts of verb lexemes: the auxiliary verb, the verb-forming particles, the participles and simple verbs are described. A, E, and irregular conjugation patterns as well as verb stem formation patterns are specified. The voice and aspect, as well as preterite (past perfect) and subjunctive verb suffixes are given. Analytical verb forms (modes) are described.

Participles and verbs forming suffixes are summarized in separate tables.

9.4 Information Appendix D- Morphemes

The numerals forming stems and the numerals names larger than a billion are given. A separate table contains the lists prefixes denoting extremely big and extremely small numbers. The types of conjunctions and modal words are summarized, as well as the patterns for word forms creation from natural sounds stems. Types of simple stems are listed and described. The categorized list of word forms components and the incompatible combinations of word-forming suffixes are summarized.

9.5 Information Appendix E- Syntax

The roles of the sentence members and the types of verb complements (objects) are listed. The word forms expressing position, direction and orientation are listed.

The alternative declension cases of Armenian, which are used in the verb phrase signatures, are described.

The types of tense, mood, aspect, voice of verbs conjugation and cases of nouns declension systems of natural languages are summarized; a brief description of each type and examples are given. Universal conjugation and declansion categories are mapped into corresponding categories of Armenian language.

9.6 Appendix F

Basics of fuzzy sets and Recurrent Neural Networks, the encoder-decoder systems are explained. The sequence-to-sequence transformation approach, which currently prevails in translation systems, is described.

10 Conclusions

Modeling of any complex system, such as a language, can be done using statistical or deterministic (algorithmic) methods. In general, the former are used when the laws of the system behavior are either unknown or very complex. In practice, the unknown is no different from the very complex. "But when the rule is too complicated, then what conforms to it is considered irregular" [Lei1686::10].

The basic grammatical laws are known and relatively simple. But they usually have many exceptions that make them complex. However, algorithmic language modeling is preferred for qualitative reasons.

The algorithmic approach to language modeling, in turn, can have two polar implementations: "brute force" (extensional) and regular or algorithmic (intensional). The former is based on the list of all the direct and inflectional word forms (lexemes), and the later - on the list of simple morphemes: stems, particles, and suffixes. In the latter case, the text forms are constructed or analyzed by algorithms that implement the grammatic rules.

The compromise solution is the most practical: to implement the simple rules in a intensional manner, and the complex - in extensional. It is obvious that when the implementation of the algorithm is much more labor-intensive than the "manual" listing of the forms produced or analyzed by it, then it makes no sense to calculate them. For example, simple stems dictionary includes complex, rarely used suffixed, alternated and other forms.

Statistical approaches, such as neural networks (machine learning, AI), are relatively simple, but require large amounts of high-quality data for training. The advantage of neural networks is in ability to solve a variety of highly complex problems in a similar way. However, the solution quality is relatively low.

It sems natural to follow these rules of thumb:

1. If the rules (for example, grammatical) are known and clear, then the algorithmic approach is preferable.
2. If not (for example, recognition of voice or writings), then statistical makes more sense.

For building a complete linguistic model, perhaps it makes sense to use statistical methods for voice and character recognition for speech coming from the outside world, and implement "deep" learning of the SN. Despite its complexity natural speech is well described by precise formulas and algorithms, because it is structural: the text parsing – the Analysis - and the text generating – the Linearization - are grammatically accurate algorithms,

11 Further studies

After productization and wider use of the proof-of-concept implementations of the suggested stemming, tagging, and lemmatization systems the outline of solutions to many of the listed below problems could be found.

1. Clarify the sequence of Armenian prefixes and find out which of them can be attached to stems and excluded from the list of suffixes.
2. Classify Armenian parts of speech using grammatical significance and roles in the sentence: a) nominal (noun and adjective; or substantive and attributive noun), b) pronoun, c) participle (without predicative, resultive, and progressive forms - these are

nouns), d) verb, e) adverb, f) adpositions, g) conjunction, as well as h) modal, h) natural sound, j) functional (negative/affirmative, and vocative/attention-grabbing) words.

3. Validate suggested patterns of common verb stem formation and Investigate their compatibility with the imperative stems formation patterns.
4. Classification of Armenian paradigmatic inflected forms:
    a. advantages and disadvantages of combining Nominal and Objective cases.
    b. categories of aspects: a) indefinite, b) dual (neutral), c) imperfect, d) perfect;
    c. categories of voice: a) active, b) neutral, c) causative, d) passive, e) reflexive.
5. Validate the significance (importance) of the classification (categorization) of verb complements per roles in the sentence.
6. Classify verbs by the verb-adverb alignment
7. Validate usefulness, completeness, and accuracy of verb signatures
8. Validate grammatical importance of detail categorization of the of the sentence member roles: agent, force, purpose, etc.
9. To study the differences of natural sounds words, pronouns, numerals and paradigms of nominal and verb forms in different languages. In particular:
    a. Categorize the unified paradigmatic types and eliminate of synonymous types.
    b. Map natural sounds words in different languages
    c. Compare the grammar Numerals naming
    d. Investigate the expressions of kinship (seems deeply studed area)
10. Map the predicate forms one-to-one into the unified conjugation grammemes.
11. Develop alternative multi-level subordination of nouns and verbs into categories and subcategories: real, abstract, animate, alterative, transitive, etc.
12. Validate considering the combination of unified declension and conjugation systems as a paradigm for lexeme.
13. Investigate the behavior of artificial neural networks during learning and functioning to identify the mapping of static and dynamic neural clusters into grammatical concepts (structures).
14. Mathematics is a subset of natural speech grammatical laws of construction and analysis for numerical, logiical, and ontological statements, according to which mathematical speech: formulas, algorithms, proofs, etc. is constructed. Separate and formalize the patententese, legalese, etc.
15. Expression of mirativity and of evidentiality in grammar of natural languages and mapping into Armenian constructs.
16. The grammar of representing of numbers (numerals) in natural languages
17. Kinship lexicon: universal and specific in natural languages
18. Grammar of direction representation in natural languages

12 References

[Bob2006] S. Bobzien. Ancient Logic. In Stanford Encyclopedia of Philosophy (online). 2006.
[Cho1975] N. Chomsky. Reflections on Language. In 'On Language'. The New Press, NY. 2007.
[Dea1998] T. Deacon. The Symbolic Species: The Co-evolution of Language and the Brain. W. W. Norton & Company. NY. 1998.
[Dev2000] K. Devlin. The Math Gene. Basic Books. 2000.

[Fre1892] Über Sinn und Bedeutung (On Sense and Reference), Zeitschrift für Philosophie und philosophische Kritik, vol. 100, 1892
[Has2006] M. Haspelmath. Terminology of case (for A. Malchukov & A. Spencer (eds.), Handbook of Case, Oxford University Press), July 2006.
[Hof1979] D.R. Hofstadter. Gödel, Escher, Bach: An Eternal Golden Braid. (20th anniversary edition) Basic Books, 1999.
[Jia2015] J. Jiang, H. Liu. Review of Lucien Tesnière, Elements of structural syntax. (John Benjamins, 2015). Journal of Linguistics · November 2015.
[Tes2015] L. Tesnière. Elements of structural syntax. Translated by Timothy Osborne and Sylvain Kahane. Amsterdam & Philadelphia, PA: John Benjamins, 2015.
[Lei1686] G.W. Leibniz. Discourse on Metaphysics. In Discourse on Metaphysics and Other Essays (Edited and Translated by Daniel Graber and Roger Ariew). Hackett Publishing Company. Indianapolis & Cambridge. 1991.
[Jah1969] G. Jahukyan. The evolution and structure of the Armenian language. Mitk, Ye., 1969.
[Jah1974] G. Jahukyan. Foundations of the theory of modern Armenian language. Publishing House of the Academy of Sciences of the Armenian SSR, Ye., 1974.

13 History of Revisions

Name	Date	Reason for Changes	Version
Aram Hayrapetyan	02/04/23	First draft.	0.1
Aram Hayrapetyan	02/20/23	Added 9.b-d in the *Further studies* section	0.2